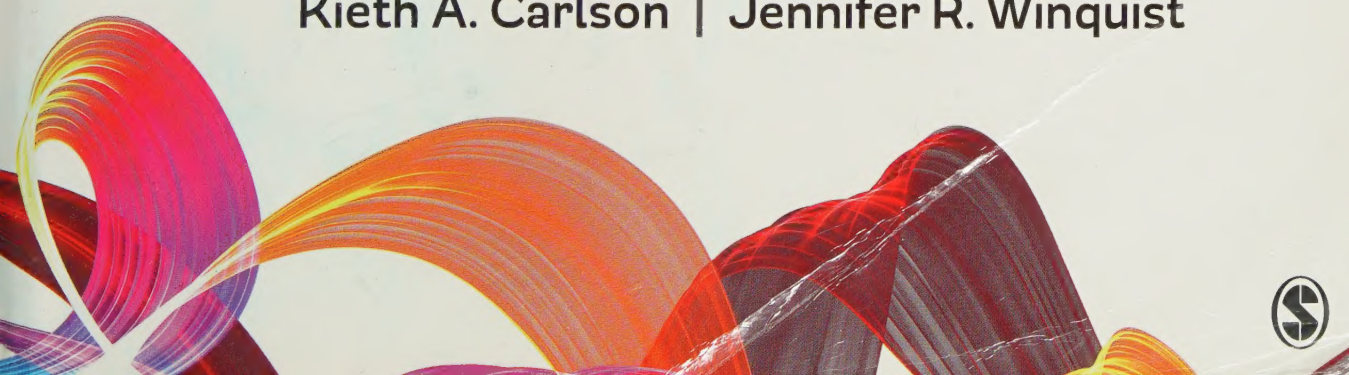


EDITION
2

AN INTRODUCTION TO STATISTICS

An Active Learning Approach

Kieth A. Carlson | Jennifer R. Winquist



	z for a Sample Mean	*Single-Sample t*	*Related t*	*Independent t*	*Correlation*
Research Situation	Testing difference between a sample mean (e.g., M = 98) and a population mean; σ known (e.g., μ = 100, σ = 15)	Testing difference between a sample mean (e.g., M = 98) and a population mean; σ unknown (e.g., μ = 100, σ = ?)	Testing difference between two related sample means (e.g., pre vs. post)	Testing difference between two sample means collected from different groups (e.g., men vs. women)	Testing relationship between two interval/ratio variables—Pearson; if either is ordinal—Spearman
1. Assumptions	-Appropriate measurement -Normality -Independence -Homogeneity of variance	-Appropriate measurement -Normality -Independence -Homogeneity of variance	-Appropriate measurement -Normality -Independence	-Appropriate measurement -Normality -Independence -Homogeneity of variance	For Pearson -Appropriate measurement -Normality -Independence -Homoscedasticity -Linear relationship
2. Hypotheses	Two-tailed H_0: μ = 100; H_1: μ ≠ 100 One-tailed H_0: μ ≤ 100; H_1: μ > 100 OR H_0: μ ≥ 100; H_1: μ < 100	Two-tailed H_0: μ = 100; H_1: μ ≠ 100 One-tailed H_0: μ ≤ 100; H_1: μ > 100 OR H_0: μ ≥ 100; H_1: μ < 100	Two-tailed H_0: $\mu_D = 0$; H_1: $\mu_D \neq 0$ One-tailed H_0: $\mu_D \leq 0$; H_1: $\mu_D > 0$ OR H_0: $\mu_D \geq 0$; H_1: $\mu_D < 0$	Two-tailed H_0: $\mu_1 = \mu_2$; H_1: $\mu_1 \neq \mu_2$ One-tailed H_0: $\mu_1 \leq \mu_2$; H_1: $\mu_1 > \mu_2$ OR H_0: $\mu_1 \geq \mu_2$; H_1: $\mu_1 < \mu_2$	Two-tailed H_0: ρ = 0; H_1: ρ ≠ 0 One-tailed H_0: ρ ≤ 0; H_1: ρ > 0 OR H_0: ρ ≥ 0; H_1: ρ < 0
3. Critical region	If two-tailed, α = .05, CV = 1.96 or −1.96 If one-tailed, α = .05, CV = 1.65 or −1.65	$df = N - 1$	$df = N - 1$	$df = (n_1 - 1) + (n_2 - 1)$	$df = N - 2$
4. Test statistic	$SEM_p = \frac{\sigma}{\sqrt{N}}$ $z = \frac{M - \mu}{SEM_p}$	$SEM_S = \frac{SD}{\sqrt{N}}$ $t = \frac{M - \mu}{SEM_S}$	$SEM_r = \frac{SD_D}{\sqrt{N}}$ $t = \frac{M_D}{SEM_r}$	$SD_p^2 = \frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{(n_1 - 1) + (n_2 - 1)}$ $SEM_i = \sqrt{\frac{SD_p^2}{n_1} + \frac{SD_p^2}{n_2}}$ $t = \frac{(M_1 - M_2)}{SEM_i}$	$SS_{XY} = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}$ $r = \frac{SS_{XY}}{\sqrt{(SS_X)(SS_Y)}}$

	z for a Sample Mean	*Single-Sample t*	*Related t*	*Independent t*	*Correlation*
5. Effect size	$d = \frac{M - \mu}{\sigma}$ .2, .5, .8	$d = \frac{M - \mu}{SD}$ .2, .5, .8	$d = \frac{M_D}{SD_D}$ .2, .5, .8	$d = \frac{M_1 - M_2}{\sqrt{SD_p^2}}$ .2, .5, .8	r^2 .01, .09, .25
6. Confidence intervals	CI for sample mean $M \pm (t_{CI})\left(\sigma/\sqrt{N}\right)$ CI for mean difference $(M - \mu) \pm (t_{CI})\left(\sigma/\sqrt{N}\right)$	CI for sample mean $M \pm (t_{CI})\left(SD/\sqrt{N}\right)$ CI for mean difference $(M - \mu) \pm (t_{CI})\left(SD/\sqrt{N}\right)$	CI for each mean $M_D \pm (t_{CI})\left(SD/\sqrt{N}\right)$ CI for mean difference $(M_1 - M_2) \pm (t_{CI})\left(SD_D/\sqrt{N}\right)$	CI for each mean $M \pm (t_{CI})\left(SD/\sqrt{N}\right)$ CI for mean difference $(M_1 - M_2) \pm (t_{CI})\left(\sqrt{SD_p^2}/\sqrt{N}\right)$	CI for Pearson $(z_r) \pm (z_{CI})\left(\frac{1}{\sqrt{N-3}}\right)$
7. Summarize	There was (or was not) a significant difference between the sample mean (*M*, *SD*) and the population mean (μ, σ), *z* (*N*) = __, *p* = __, *d* = __, 95% CI [LB, UB]. *If appropriate, indicate which mean was significantly higher and describe the effect size.*	There was (or was not) a significant difference between the sample mean (*M*, *SD*) and the population mean (μ), *t* (*df*) = __, *p* = __, *d* = __, 95% CI [*LB, UB*]. *If appropriate, indicate which mean was significantly higher and describe the effect size.*	There was (or was not) a significant difference between the pre-treatment sample mean (*M*, *SD*) and the post treatment sample mean (*M*, *SD*), *t* (*df*) = __, *p* = __, *d* = __, 95% CI [LB, UB]. *If appropriate, indicate which mean was significantly higher and describe the effect size.*	There was (or was not) a significant difference between the Sample 1 mean (*M*, *SD*) and the Sample 2 mean (*M*, *SD*), *t* (*df*) = __ , *p* = __, *d* = __, 95% CI [LB, UB]. *If appropriate, indicate which mean was significantly higher and describe the effect size.*	There was (or was not) a linear association between Variable 1 and Variable 2, *r* (*df*) = ___, *p* = ___, 95% CI [LB, UB].
8. SPSS instructions for significance test	Not available	-Analyze -Compare Means -One-Sample *t* Test -Move DV into the Test Variables box -Change Test Value to μ -Click OK	-Analyze -Compare Means -Paired-Samples T Test -Move both IV conditions into Paired Variables box -Click OK	-Analyze -Compare Means -Independent-Samples T Test -Move IV into Grouping Variable box -Click Define Groups -Enter values that designate each IV condition -Move DV into Test Variables box -Click OK	For scatterplot: -Graph, Legacy Dialogs, Scatter/Dot, -Simple scatter -Click Define -Place variables on x- and y-axes For test: -Analyze, Correlate, Bivariate -Move variables into Variables box -Select Pearson or Spearman. Click OK

An Introduction to Statistics

Second Edition

Sara Miller McCune founded SAGE Publishing in 1965 to support the dissemination of usable knowledge and educate a global community. SAGE publishes more than 1000 journals and over 800 new books each year, spanning a wide range of subject areas. Our growing selection of library products includes archives, data, case studies and video. SAGE remains majority owned by our founder and after her lifetime will become owned by a charitable trust that secures the company's continued independence.

Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

An Introduction to Statistics

AN ACTIVE LEARNING APPROACH

Second Edition

Kieth A. Carlson

Jennifer R. Winquist

Valparaiso University

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC | Melbourne

SAGE

FOR INFORMATION:

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
1 Oliver's Yard
55 City Road
London, EC1Y 1SP
United Kingdom

SAGE Publications India Pvt. Ltd.
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Asia-Pacific Pte. Ltd.
3 Church Street
#10–04 Samsung Hub
Singapore 049483

Copyright © 2018 by SAGE Publications, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

All trademarks depicted within this book, including trademarks appearing as part of a screenshot, figure, or other image are included solely for the purpose of illustration and are the property of their respective holders. The use of the trademarks in no way indicates any relationship with, or endorsement by, the holders of said trademarks. SPSS is a registered trademark of International Business Machines Corporation.

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Names: Carlson, Kieth A., author. | Winquist, Jennifer R., author.

Title: Introduction to statistics : an active learning approach / Kieth A. Carlson & Jennifer R.Winquist, Valparaiso University.

Description: Second edition. | Los Angeles : SAGE, [2018] | Includes bibliographical references and index.

Identifiers: LCCN 2016039108 | ISBN 9781483378732 (pbk. : alk. paper)

Subjects: LCSH: Social sciences—Statistical methods. | Statistics.

Classification: LCC HA29 .C288 2018 | DDC 519.5—dc23 LC record available at https://lccn.loc.gov/2016039108

This book is printed on acid-free paper.

SUSTAINABLE FORESTRY INITIATIVE
Certified Chain of Custody
Promoting Sustainable Forestry
www.sfiprogram.org
SFI-01268
SFI label applies to text stock

17 18 19 20 21 10 9 8 7 6 5 4 3 2

Acquisitions Editor: Abbie Rickard
Editorial Assistant: Alexander Helmintoller
eLearning Editor: Morgan Shannon
Production Editor: Kelly DeRosa
Copy Editor: Gillian Dickens
Typesetter: C&M Digitals (P) Ltd.
Proofreader: Jeanne Busemeyer
Indexer: Wendy Jo Dymond
Cover Designer: Rose Storey
Marketing Manager: Katherine Hepburn

Contents

Preface	xix
About the Authors	xxv
Chapter 1. Introduction to Statistics and Frequency Distributions	1
Chapter 2. Central Tendency	39
Chapter 3. Variability	65
Chapter 4. *z* Scores	95
Chapter 5. The Distribution of Sample Means and *z* for a Sample Mean	113
Chapter 6. Hypothesis Testing With *z* Scores	145
Chapter 7. Single-Sample *t* Test	207
Chapter 8. Estimation With Confidence Intervals	241
Chapter 9. Related Samples *t* Test	271
Chapter 10. Independent Samples *t* Test	315
Chapter 11. One-way Independent Samples ANOVA	367
Chapter 12. Two-Factor ANOVA or Two-Way ANOVA	439
Chapter 13. Correlation and Regression	513
Chapter 14. Goodness-of-Fit and Independence Chi-Square Statistics	567
Appendices	597
Index	619



Digitized by the Internet Archive
in 2023 with funding from
Kahle/Austin Foundation

https://archive.org/details/introductiontost0000carl_e7n9

Detailed Contents

Preface xix

About the Authors xxv

1 Introduction to Statistics and Frequency Distributions 1

How to Be Successful in This Course 1
Math Skills Required in This Course 3
Why Do You Have to Take Statistics? 4
Statistics and the Helping Professions 4
Hypothesis Testing, Effect Size, and Confidence Intervals 5
Testing Causal Hypotheses 6
Populations and Samples 7
Independent and Dependent Variables 8
Scales of Measurement 9
Discrete Versus Continuous Variables 12
Graphing Data 12
Shapes of Distributions 15
Frequency Distribution Tables 19
SPSS 20
Overview of the Activity 24
Activity 1.1: Frequency Distributions 24
Chapter 1 Practice Test 32

2 Central Tendency 39

Central Tendency 39
Computing the Mean 42
Find the Median 45
Find the Mode 47
SPSS 47
Overview of the Activity 51
Activity 2.1: Central Tendency 51
Chapter 2 Practice Test 61

3 Variability 65

Population Variability 65
Steps in Computing a Population's Standard Deviation 66
Step 1: Compute the Deviation Scores ($X - \mu$) 66
Step 2: Square the Deviation Scores $(X - \mu)^2$ 67
Step 3: Compute the Sum of the Squared Deviation Scores, $SS = \Sigma(X - \mu)^2$ 68
Step 4: Compute the Variance (σ^2) 69
Step 5: Compute the Standard Deviation (σ) 69
Sample Variability 72
Steps 1 Through 3: Obtaining the SS 73
Step 4: Compute the Sample Variance (SD^2) 73
Step 5: Compute the Sample Standard Deviation (SD) 74
SPSS 75
Overview of the Activity 78

Activity 3.1: Variability 78

Chapter 3 Practice Test 91

4 *z* Scores 95

z for a Single Score 95
Computing a z for an Individual Score 96
Interpreting the z for a Single Score 96
Using *X* to Find Important "Cut Lines" 97
z Scores and the Standard Normal Curve 98
Example 1: Positive *z* Score 100
Compute the z Score 100
Draw a Normal Distribution, and Shade the Area You Are Interested In 101
Use a Unit Normal Table (Located in Appendix A of This Book) to Find the Area of the Shaded Proportion of the Curve 101
Example 2: Negative *z* Score 102
Draw a Normal Distribution, and Shade the Area You Are Interested In 102
Use a Unit Normal Table to Find the Area That Is Shaded 102
Example 3: Proportion Between Two *z* Scores 103
Draw a Normal Distribution, and Shade the Area You Are Interested In 103
Use a Unit Normal Table to Find the Area That Is Shaded 103
Overview of the Activity 104

Activity 4.1: *z* Scores and Probabilities 104

Chapter 4 Practice Test 111

5 The Distribution of Sample Means and *z* for a Sample Mean 113

Sampling and Sampling Error 113
Distribution of Sample Means 115

z for a Sample Mean 122
Example: Computing and Interpreting the *z* for a Sample Mean 124
Step 1: Compute the Observed Deviation 124
Step 2: Compute the Deviation Expected by Sampling Error 124
Step 3: Compute the Ratio Between Observed and Expected Deviation (z for a Sample Mean) 124
Step 4: Locate the z Score in the Distribution 124
Step 5: Look Up the z Score 125
Step 6: Interpret the z Score 125
Exact Probabilities Versus Probability Estimates 125
Overview of the Activities 126

Activity 5.1: Introduction to Distributions of Sample Means 126

Activity 5.2: Central Limit Theorem 136
Chapter 5 Practice Test 142

6 Hypothesis Testing With *z* Scores 145

Introduction to Hypothesis Testing 145
Hypothesis Testing With *z* for a Sample Mean Example (One-Tailed) 146
Step 1: Examine Variables to Assess Statistical Assumptions 146
Step 2: State the Null and Research Hypotheses Symbolically and Verbally 148
Step 3: Define the Critical Region 150
Step 4: Compute the Test Statistic (z for a Sample Mean) 154
Step 5: Compute the Effect Size, and Describe It as Small, Medium, or Large 155
Step 6: Interpreting the Results of the Hypothesis Test Using a z for a Sample Mean 157
What Does It Mean to Describe Something as "Statistically Significant"? 157
Errors in Hypothesis Testing 158
Hypothesis Testing Rules 160
What Is a *p* Value? 162
Why Statisticians "Fail to Reject the Null" Rather Than "Accept the Null" 164
Why Scientists Say "This Research Suggests" Rather Than "This Research Proves" 165
Overview of the Activities 166

Activity 6.1: Hypothesis Testing 166

Activity 6.2: Critical Values, *p* Values, and the Null Hypothesis 177

Activity 6.3: Statistical Power, Type I Error, and Type II Error 182

Activity 6.4: Hypothesis Testing and Effect Size 195
Chapter 6 Practice Test 204

7 Single-Sample *t* Test 207

Single-Sample *t* Test 207
Conceptual Information 208
One-Tailed Single-Sample *t* Test Example 211
Step 1: Examine the Statistical Assumptions 211

Step 2: State the Null and Research Hypotheses Symbolically and Verbally 212
Step 3: Use Sample Size to Compute Degrees of Freedom and Define the Critical Region 213
Step 4: Compute the Test Statistic (Single-Sample t Test) 214
Step 5: Compute an Effect Size and Describe It 216
Step 6: Interpreting the Results of the Hypothesis Test 216
Two-Tailed Single-Sample *t* Test Example 217
Step 1: Examine the Statistical Assumptions 219
Step 2: State the Null and Research Hypotheses Symbolically and Verbally 219
Step 3: Use Sample Size to Compute Degrees of Freedom and Define the Critical Regions 220
Step 4: Compute the Test Statistic (Single-Sample t Test) 220
Step 5: Compute an Effect Size and Describe It 222
Step 6: Interpreting the Results of the Hypothesis Test 222
Other Alpha Levels 222
SPSS 223
Overview of the Activity 226
Activity 7.1: Single-Sample *t* Test 227
Chapter 7 Practice Test 234

8 Estimation With Confidence Intervals 241

Three Statistical Procedures With Three Distinct Purposes 241
Logic of Confidence Intervals 244
Computing a Confidence Interval for a Population Mean 245
Computing Confidence Intervals for a Mean Difference 247
Reporting Confidence Intervals in APA Style 249
Confidence Intervals for Effect Sizes 249
Interpretations of Confidence Intervals 250
SPSS 251
Overview of the Activity 252
Activity 8.1: Estimating Sample Means and Sample Mean differences 253
Chapter 8 Practice Test 267

9 Related Samples *t* Test 271

Repeated/Related Samples *t* Test 271
Logic of the Single-Sample and Repeated/Related Samples *t* Tests 273
Related Samples *t* (Two-Tailed) Example 274
Step 1: Examine the Statistical Assumptions 274
Step 2: State the Null and Research Hypotheses Symbolically and Verbally 275
Step 3: Compute the Degrees of Freedom and Define the Critical Region 276
Step 4: Compute the Test Statistic (Related Samples t) 277
Step 5: Compute an Effect Size and Describe It 279

Step 6: Interpreting the Results of the Hypothesis Test 279
Related Samples *t* (One-Tailed) Example 280
Step 1: Examine the Statistical Assumptions 280
Step 2: State the Null and Research Hypotheses Symbolically and Verbally 280
Step 3: Compute the Degrees of Freedom and Define the Critical Region 281
Step 4: Compute the Test Statistic (Related Samples t) 281
Step 5: Compute an Effect Size and Describe It 283
Step 6: Interpreting the Results of the Hypothesis Test 283
Statistical Results, Experimental Design, and Scientific Conclusions 284
SPSS 284
Overview of the Activities 289
Activity 9.1: Hypothesis Testing with the Related Samples *t* Test (or Dependent *t* Test) 289
Activity 9.2: Combining Significance Testing, Effect Sizes, and Confidence Intervals 299
Chapter 9 Practice Test 309

10 Independent Samples *t* Test 315

Independent Samples *t* 315
Conceptual Formula for the Independent Samples *t* 319
Two-Tailed Independent *t* Test Example 320
Step 1: Examine the Statistical Assumptions 320
Step 2: State the Null and Research Hypotheses Symbolically and Verbally 321
Step 3: Compute the Degrees of Freedom and Define the Critical Region 322
Step 4: Compute the Test Statistic 322
Step 5: Compute an Effect Size and Describe It 325
Step 6: Interpreting the Results of the Hypothesis Test 326
One-Tailed Independent *t* Test Example 326
Step 1: Examine the Statistical Assumptions 326
Step 2: State the Null and Research Hypotheses Symbolically and Verbally 326
Step 3: Compute the Degrees of Freedom and Define the Critical Region 327
Step 4: Compute the Test Statistic 328
Step 5: Compute an Effect Size and Describe It 329
Step 6: Interpreting the Results of the Hypothesis Test 329
Other Alpha Levels 330
SPSS 330
Overview of the Activities 336
Activity 10.1: Hypothesis Testing With the Independent *t* Test 336
Activity 10.2: A Two-Tailed Independent *t* Test 347
Activity 10.3: How to Choose the Correct Statistic 353
Activity 10.4: Comparing Independent, Matched, and Repeated Research Designs 355
Activity 10.5: Confidence Intervals for Mean Differences Between Independent Samples 357
Chapter 10 Practice Test 362

11 One-Way Independent Samples ANOVA 367

Independent Samples ANOVA 367
Other Names 368
Logic of the ANOVA 368
An Example ANOVA Problem 371
Step 1: Examine Variables to Assess Statistical Assumptions 372
Step 2: State the Null and Research Hypotheses 373
Step 3: Define the Critical Value of F 374
Step 4: Computing the Test Statistic (Independent ANOVA) 374
Step 5: Compute the Effect Size and Describe It 380
Step 6: Summarize the Results 382
An Additional Note on ANOVAs: Family-Wise Error and Alpha Inflation 382
SPSS 383
Overview of the Activities 388

Activity 11.1: Computing One-Way Independent ANOVAs 388

Activity 11.2: Computing One-Way Independent ANOVAs in SPSS 398

Activity 11.3: Independent ANOVA With SPSS 400

Activity 11.4: Understanding Within- and Between-Group Variability 411

Activity 11.5: Confidence Intervals 423

Activity 11.6: Choose the Correct Statistic 428
Chapter 11 Practice Test 431

12 Two-Factor ANOVA or Two-Way ANOVA 439

Purpose of the Two-Way ANOVA 439
Describing Factorial Designs 440
Logic of the Two-Way ANOVA 441
Example of a Two-Way ANOVA 445
Step 1: Examine Variables to Assess Statistical Assumptions 445
Step 2: Set Up the Null and Research Hypotheses 446
Step 3: Define the Critical Region 450
Step 4: Compute the Test Statistics (Three F Tests) 452
Step 5: Compute the Effect Sizes 454
Step 6: Writing Up the Results of a Two-Way ANOVA 457
SPSS 457
Overview of the Activities 464

Activity 12.1: Two-Factor ANOVAs I 464

Activity 12.2: Two-Factor ANOVAs II 476

Activity 12.3: Two-Factor ANOVAs III 481

Activity 12.4: One-Way and Two-Way ANOVA Review 490

Activity 12.5: Choose the Correct Statistic 502
Chapter 12 Practice Test 506

13 Correlation and Regression 513

When to Use Correlations and What They Can Tell You 513
Review of *z* Scores 514
The Logic of Correlation 515
Direction and Strength of Correlation Coefficients 516
Computational Formulas 519
Spearman's (r_s) Correlations 521
Using Scatterplots Prior to Correlation Coefficients 521
Alternative Use for Correlation 523
Correlation and Causation 523
Hypothesis Testing With Correlation 524
Two-Tailed Pearson's Correlation Example 525
Step 1: Assess Statistical Assumptions 526
Step 2: State the Null and Research Hypotheses Symbolically and Verbally 527
Step 3: Define the Critical Region 527
Step 4: Compute the Test Statistic (Pearson's r) 528
Step 5: Compute the Effect Size (r^2) and Describe It 528
Step 6: Summarize the Results 529
One-Tailed Pearson's Correlation Example 529
Step 1: Assess Statistical Assumptions 529
Step 2: State the Null and Research Hypotheses Symbolically and Verbally 529
Step 3: Define the Critical Region 531
Step 4: Compute the Test Statistic (Pearson's r) 531
Step 5: Compute the Effect Size (r^2) and Describe It 531
Step 6: Summarize the Results 531
What If You Need to Do a Spearman's Correlation? 532
Confidence Intervals 532
SPSS 533
Overview of the Activities 536
Activity 13.1: Correlations 536
Activity 13.2: Confidence Intervals for Correlations 546
Activity 13.3: Spearman's Correlation 550
Activity 13.4: Introduction to Regression and Prediction 552
Activity 13.5: Choose the Correct Statistic 560
Chapter 13 Practice Test 563

14 Goodness of Fit and Independence Chi-Square Statistics 567

Overview of Chi-Square 567
Logic of the Chi-Square Test 569
Comparing the Goodness-of-Fit Chi-Square and the Chi-Square for Independence 570
Goodness-of-Fit Chi-Square Example 571
Step 1: Examine Statistical Assumptions 571
Step 2: State the Null and Research Hypotheses 571
Step 3: Compute the df *and Define the Critical Region 571*
Step 4: Compute the Test Statistic (Goodness-of-Fit Chi-Square) 572
Step 5: Interpret the Results 573
Chi-Square for Independence 573
Step 1: Examine Statistical Assumptions 573
Step 2: State the Null and Research Hypotheses 574
Step 3: Compute the df *and Define the Critical Region 574*
Step 4: Compute the Test Statistic (Chi-Square Test for Independence) 574
Step 5: Compute the Effect Size and Interpret It as Small, Medium, or Large 576
Step 6: Interpret the Results 577
SPSS 578
Overview of the Activities 583
Activity 14.1: Goodness-of-fit chi-square and chi-square for independence 583
Activity 14.2: Choose the Correct Statistic 588
Chapter 14 Practice Test 593

Appendices 597

Appendix A
Unit Normal Table (*z* Table) 597
Appendix B
One-Tailed Probabilities *t* Table 601
Two-Tailed Probabilities *t* Table 602
Appendix C
F Table ($\alpha = .05$) 603
F Table ($\alpha = .01$) 605
Appendix D
The Studentized Range Statistic (*q*) Table 607
Appendix E
One-Tailed Pearson's Correlation Table 609
Two-Tailed Pearson's Correlation Table 610
Appendix F
Spearman's Correlation Table 611

Appendix G
Fisher *r* to *z* Table 612
Appendix H
Critical Values for Chi-Square 613
Appendix I
Computing *SSs* for Factorial ANOVA 614
Appendix J
Choosing Correct Test Statistics 616

Index 619

Preface

THE STORY OF THIS TEXT

Several years ago, we attended a teaching workshop in which the speaker described a common experience in college classrooms and the pedagogical problems it frequently creates. Instructors carefully define basic concepts (e.g., population, sample) and gradually progress to applying those concepts to more complex topics (e.g., sampling error) as the end of class approaches. Then students attempt homework assignments covering the more complicated topics. All too frequently, students think they understand things while listening to us in class, but when they attempt homework on their own, they have difficulty. While some students can eventually figure things out, others become frustrated; still others give up. The teaching workshop made us recognize, reluctantly, this happened to us (and our students) in our statistics classes. While we did our best to address this problem by refining our lectures, our students still struggled with homework assignments, and we were disappointed with their exam performance. Students frequently said to us, "I understand it when you do it in class, but when I try it on my own it doesn't make sense." This common experience motivated us to change our stats classes and, eventually, to write the first edition of this text.

We decided that we needed to change our course so that

1. students came to class understanding basic concepts and
2. students had an opportunity to *use* challenging concepts in class when we were there to answer their questions immediately,
3. students started to interpret and report statistical results like researchers.

We started by emphasizing the importance of actually reading the text before class. Even though we were using excellent statistics texts, many students insisted that they needed lectures to help them understand the text. Eventually, we opted for creating our own readings that emphasize the basics (i.e., the "easy" stuff). We embedded relatively easy reading questions to help students *read with purpose* so they came to class understanding the basic concepts. Next, over several years, we developed activities that reinforced the basics as well as introduced more challenging material (i.e., the "hard stuff"). Hundreds of students completed these challenging activities in our courses. After each semester, we strove to improve every activity based on our students' feedback and exam performance.

Our statistics courses are dramatically different from what they were a decade ago. In our old classes, few students read prior to class, and most class time was spent lecturing on the material in the book. In our current stats courses, students answer online reading questions prior to class, we give very brief lectures at the beginning of class, and students complete activities (i.e., assignments) during class.

We've compared our current students' attitudes about statistics to those taking our more traditional statistics course (Carlson & Winquist, 2011) and found our current students to be more confident in their ability to perform statistics and to like statistics more than their peers. We've also learned that after completing this revised statistics course, students score nearly a half a standard deviation higher on a nationally standardized statistics test that they take during their senior year (approximately 20 months after taking the course) compared to students taking the more traditional course (Winquist & Carlson, 2014).

Of course, not all our students master the course material. Student motivation still plays an important part in student learning. If students don't do the reading or don't work on understanding the assignments in each chapter, they will still struggle. In our current courses, we try to create a class that encourages students to read and complete the assignments by giving points for completing them. We have found that, if students do these things, they do well in our courses. We have far fewer struggling students in our current courses than we had in our traditional course, even though our exams are more challenging.

WHAT IS NEW IN THE SECOND EDITION

If you used the first edition of the text, the first thing you might notice is that the second edition has 14 chapters rather than 16, but the text is actually longer. In the first edition, all hypothesis tests followed the same five steps and statistical assumptions were addressed in Chapter 16. In the second edition, we eliminated Chapter 16 and included assessing the statistical assumptions as the first step of a six-step hypothesis-testing process. While talking about the statistical assumptions within every chapter is less concise, this repetition helps students recognize that different statistical tests analyze different types of variables. In addition, in response to reviewers' comments, we also combined Chapters 6 and 7 from the first edition into a single chapter in the second edition. Finally, in the first edition, we introduced the basics in the chapter and then added more complex material in the activities. Although this simplified the readings for students, it also made the book harder for students to use as a reference. In this edition, we include the more complex material in the chapters but kept the reading questions relatively simple. This way, students are exposed to the material prior to working with the more complex ideas in the assignments. Reflecting the rising prominence of confidence intervals in contemporary research and the most recent APA publication manual, we greatly expanded our coverage of confidence intervals in the second edition. We added integrative assignments in the related *t*, independent *t*, one-way analysis of variance (ANOVA), and correlation chapters to reinforce the different information researchers obtain from significance tests, effect sizes, and confidence intervals. These assignments encourage students to do more than "crunch numbers" by asking them to think like researchers, integrating information from significance tests, effect sizes, and confidence intervals.

Other noteworthy changes to the second edition include the following:

- New assignments are included on the hand calculations of a one-way ANOVA, running one-way ANOVA in SPSS, the differences between one-way and two-way ANOVA, and Spearman correlation.

- Twelve of the 14 chapters have been rewritten using more interesting examples from psychological research.
- Assignments contain fewer open-ended questions so students can check their own answers more accurately.
- Added coverage of effect sizes for pairwise comparisons.
- Added practice tests at the end of each chapter.

HOW TO USE THIS BOOK

This text certainly could be used in a lecture-based course in which the activities function as detailed, conceptually rich homework assignments. We also are confident that there are creative instructors and students who will find ways to use this text that we never considered. However, it may be helpful to know how we use this text. In our courses, students read the chapters and answer online reading questions prior to class. We allow them to retake the reading questions to correct any errors prior to class for half of the points they missed. We begin classes with brief lectures (about 15 minutes), and then students work for the remaining 60 minutes to complete the assignments during class. There are a number of advantages to this approach. One advantage is that students do the easier work (i.e., answering foundational questions) outside of class and complete the more difficult work in class when peers and an instructor can answer their questions. Another advantage is that students work at their own paces. We have used this approach for several years with positive results (Carlson & Winquist, 2011; Winquist & Carlson, 2014).

This approach encourages students to review and correct misunderstandings on the reading questions as well as the assignments. Mistakes are inevitable and even desirable. After all, each mistake is an opportunity to learn. In our view, students should first engage with the material without concern about evaluation. Therefore, we provide the final answers to all assignments to our students. Students then focus on finding their answers, checking them, and then correcting mistakes. We collect their answers to confirm that they showed how they arrived at each answer. We give points for completion (and showing work). Over the years, these assignment points have constituted between 7% and 17% of students' course grades. A simpler option we tried is telling students that completing the activities is essential to success in the course and not confirm activity completion at all. When we did this, we found greater variability in activity completion and exam performance.

UNIQUE FEATURES OF THIS TEXT

By now you probably recognize that this is not a typical statistics text. For ease of review, we've listed and described the two most unique aspects of this text:

- *Embedded reading questions*—All 14 chapters contain embedded reading questions that focus students' attention on the key concepts *as they read* each paragraph/section of the text. Researchers studying reading comprehension report that similar embedded questions help students with lower reading abilities achieve levels of performance comparable to that of students with greater reading abilities (Callender & McDaniel, 2007).

- *Activity (Assignment) sections*—All 14 chapters contain active learning assignments, called *Activities*. While the 14 chapters start by introducing foundational concepts, they are followed by activity sections in which students *test or demonstrate their understanding of basic concepts while they read detailed explanations of more complex statistical concepts*. When using most traditional textbooks, students perform statistical procedures *after* reading multiple pages. This text adopts a workbook approach in which students are actively performing tasks *while* they read explanations. Most of the activities are self-correcting, so if students misunderstand a concept, it is corrected early in the learning process. After completing these activities, students are far more likely to understand the material than when they simply read the material.

OTHER HELPFUL FEATURES

- *Learning objectives*—Each chapter and activity begin with clear learning objectives.
- *Practice tests*—All 14 chapters conclude with a practice test for solidifying student learning.
- *IBM® SPSS® Statistics**—All chapters contain detailed step-by-step instructions for conducting statistical procedures with SPSS as well as annotated explanations of SPSS output.
- *Emphasis on understanding*—Chapters use definitional formulas to explain the logic behind each statistical procedure and rely on SPSS for more advanced computations (e.g., factorial ANOVAs).
- *Writing results in APA format*—Many activity questions highlight how to write about statistical analyses in scholarly ways.

ANCILLARIES

- *Instructors' manual*—Includes lecture outlines and detailed answers to activities.
- *Blackboard cartridges*—Includes reading questions, practice tests, self-test questions, and activity answers.
- *Empirically validated test bank questions*—Exam questions that we used in our classes are available to instructors of the course.
- *Self-examination questions*—Additional sample examination questions are available to students on the Sage Publications website.
- *Short PowerPoint slideshows for most Activities*.

APPROPRIATE COURSES

This text is ideal for introductory statistics courses in psychology, sociology, social work, and the health, exercise, or life sciences. The text would work well for any course intending to teach the statistical procedures of hypothesis testing, effect sizes, and confidence intervals that are commonly used in the behavioral sciences.

*SPSS is a registered trademark of International Business Machines Corporation.

ACKNOWLEDGMENTS

We thank Barbara E. Walvoord for inspiring us to write this text. We thank the many reviewers (listed below) who helped us improve the text with their insightful critiques and comments.

Elizabeth Axel, Adelphi University

Ray Garza, Texas A&M International University

Carolyn J. Mebert, University of New Hampshire

Lyon Rathbun, University of Texas, Brownsville

T. Siva Tian, University of Houston

We greatly appreciate the invaluable feedback of our students, without whom this text would not have been possible.

Finally, we are grateful to SAGE Publications for giving us the opportunity to share this text with others.

REFERENCES

Callender, A. A., & McDaniel, M. A. (2007). The benefits of embedded question adjuncts for low and high structure builders. *Journal of Educational Psychology, 99*(2), 339–348.

Carlson, K. A., & Winquist J. R. (2011). Evaluating an active learning approach to teaching introductory statistics: A classroom workbook approach. *Journal of Statistics Education, 19*(1). Retrieved from http://www.amstat.org/publications/jse/v19n1/carlson.pdf

Winquist, J. R., & Carlson, K. A. (2014). Flipped statistics class results: Better performance than lecture over on year later. *Journal of Statistics Education, 22*(3). Retrieved from http://www.amstat.org/publications/jse/v22n3/winquist.pdf

About the Authors

Kieth A. Carlson received his PhD in Experimental Psychology with an emphasis in Cognitive Psychology from the University of Nebraska in 1996. He is currently Professor of Psychology at Valparaiso University. He has published research on visual attention, memory, student cynicism toward college, and active learning. He enjoys teaching a wide range of courses including statistics, research methods, sensation and perception, cognitive psychology, learning psychology, the philosophy of science, and the history of psychology. Dr. Carlson was twice honored with the Teaching Excellence Award from the United States Air Force Academy.

Jennifer R. Winquist is currently Professor of Psychology at Valparaiso University. Dr. Winquist received her PhD in Social Psychology from the University of Illinois at Chicago and her bachelor's degree in Psychology from Purdue University. She has published research on self-focused attention, group decision making, distributive justice, and the scholarship of teaching and learning. Dr. Winquist regularly teaches courses in introductory and advanced statistics and research methods.

CHAPTER 1

Introduction to Statistics and Frequency Distributions

LEARNING OBJECTIVES

After reading this chapter, you should be able to do the following:

- Explain how you can be successful in this course
- Use common statistical terms correctly in a statistical context
  - Statistic, parameter, sample, population, descriptive statistics, inferential statistics, sampling error, and hypothesis testing
- Identify the scale of measurement of a variable (nominal, ordinal, or interval/ratio)
- Determine if a variable is discrete or continuous
- Create and interpret frequency distribution tables, bar graphs, histograms, and line graphs
- Explain when to use a bar graph, histogram, and line graph
- Enter data into SPSS and generate frequency distribution tables and graphs

HOW TO BE SUCCESSFUL IN THIS COURSE

Have you ever read a few pages of a textbook and realized you were not thinking about what you were reading? Your mind wandered to topics completely unrelated to the text, and you could not identify the point of the paragraph (or sentence) you just read. For most of us, this experience is not uncommon even when reading books that we've chosen to read for pleasure. Therefore, it is not surprising that our minds wander while reading textbooks. Although this lack of focus is understandable, it seriously hinders effective reading. Thus, one goal of this book is to discourage mind wandering and to encourage *reading with purpose*. To some extent, you need to force yourself to read with purpose. As you read each paragraph, ask, "What is the purpose of this paragraph?" or "What am I supposed to learn from this paragraph?"

Reading Question

1. Reading with purpose means
   a. thinking about other things while you are reading a textbook.
   b. actively trying to extract information from a text by focusing on the main point of each paragraph.

This text is structured to make it easier for you to read with purpose. The chapters have frequent reading questions embedded in the text that make it easier for you to remember key points from preceding paragraphs. Resist the temptation to go immediately to the reading questions and search for answers in the preceding paragraphs. *Read first, and then answer the questions as you come to them.* Using this approach will increase your memory for the material in this text.

Reading Question

2. Is it better to read the paragraph and then answer the reading question or to read the reading question and then search for the answer? It's better to
   a. read the paragraph, then answer the reading question.
   b. read the reading question, then search for the question's answer.

After reading the chapters, you should have a basic understanding of the material that will provide the foundation you need to work with the more complex material in the activities. When completing these activities, you will demonstrate your understanding of basic material from the reading (by answering questions) before you learn more advanced topics. Your emphasis when working on the activities should be on understanding why the answers are correct. If you generate a wrong answer, figure out your error. We often think of errors as things that should be avoided at all costs. However, quite the opposite is true. Making mistakes and fixing them is how you learn. Every error is an opportunity to learn. If you find your errors and correct them, you will probably not repeat the error. Resist the temptation to "get the right answer quickly." It is more important that you understand why every answer is correct.

Reading Question

3. Which of the following best describes the activities in this book?
   a. The material in the activities is more complex than the material in the readings.
   b. All of the new material is in the reading. The activities are simply meant to give you practice with the material in the reading.

Reading Question

4. When completing activities, your primary goal should be to get the correct answer quickly.
   a. True
   b. False

At the end of each chapter, there are is a "Practice Test." After you complete the assigned activities in a chapter (and you understand why every answer is correct), you should complete the practice test. Most students benefit from a few repetitions of each problem type. The additional practice helps consolidate what you have learned so you don't forget it during tests. Finally, use the activities and the practice tests to study. Then, *after* you understand all of the activities and all of the practice tests, assess your understanding by taking an additional self-test on the SAGE website. Try to duplicate a testing situation as much as possible. Just sit down with a calculator and have a go at it. If you can do the self-test, you should feel confident in your knowledge of the material. Taking practice tests days before your actual test will give you time to review material if you discover you did not understand something.

Testing yourself is also a good way to lessen the anxiety that can occur during testing. Again, additional practice test questions are available on the SAGE website.

Reading Question

5. How should you use the self-tests?
   a. Use them to study; complete them open-book so you can be sure to look up all the answers.
   b. Use them to test what you know days before the exam; try to duplicate the testing situation as much as possible.

MATH SKILLS REQUIRED IN THIS COURSE

Students often approach their first statistics course with some anxiety. The primary source of this anxiety seems to be a general math anxiety. The good news is that the math skills required in this course are fairly basic. You need to be able to add, subtract, multiply, divide, square numbers, and take the square root of numbers using a calculator. You also need to be able to do some basic algebra. For example, you should be able to solve the following equation for X: $22 = \frac{X}{3}$. [The correct answer is $X = 66$.]

Reading Question

6. This course requires basic algebra.
   a. True
   b. False

Reading Question

7. Solve the following equation for X: $30 = \frac{X}{3}$. 30 × 3 = 90
   a. 10
   b. 90

You will also need to follow the correct order of mathematical operations. As a review, the correct order of operations is (1) the operations in parentheses, (2) exponents, (3) multiplication or division, and (4) addition or subtraction. Some of you may have learned the mnemonic, *P*lease *E*xcuse *M*y *D*ear *A*unt *S*ally, to help remember the correct order. For example, when solving the following equation, $(3 + 4)^2$, you would first add (3 + 4) to get 7 and then square the 7 to get 49. Try to solve the next more complicated problem. The answer is 7.125. If you have trouble with this problem, talk with your instructor about how to review the necessary material for this course.

$$X = \frac{(6-1)3^2 + (4-1)2^2}{(6-1) + (4-1)}.$$

Reading Question

8. Solve the following equation for X: $X = \frac{(3-1)4^2 + (5-1)3^2}{(3-1) + (5-1)}$. (2)16 (4)9 → 32 + 36 = 68 = 11.3
   a. 11.33
   b. 15.25

You will be using a calculator to perform computations in this course. You should be aware that order of operations is very important when using your calculator. Unless you are very comfortable with the parentheses buttons on your calculator, we recommend that you do one step at a time rather than trying to enter the entire equation into your calculator.

Reading Question

9. Order of operations is only important when doing computations by hand, not when using your calculator.

a. True

b. False

Although the math in this course should not be new, you may see new notation throughout the course. When you encounter new notation, relax and realize that the notation is simply a shorthand way of giving instructions. While you will be learning how to *interpret* numbers in new ways, the actual mathematical skills in this course are no more complex than the order of operations. The primary goal of this course is teaching you to use numbers to make decisions. Occasionally, we will give you numbers solely to practice computation, but most of the time you will use the numbers you compute to make decisions within a specific, real-world context.

WHY DO YOU HAVE TO TAKE STATISTICS?

You are probably reading this book because you are required to take a statistics course to complete your degree. Students majoring in business, economics, nursing, political science, premedicine, psychology, social work, and sociology are often required to take at least one statistics course. There are a lot of different reasons why statistics is a mandatory course for students in these varied disciplines. The primary reason is that in every one of these disciplines, people make decisions that have the potential to improve people's lives, and these decisions should be informed by data. For example, a psychologist may conduct a study to determine if a new treatment reduces the symptoms of depression. Based on this study, the researcher will need to decide if the treatment is effective or not. If the wrong decision is made, an opportunity to help people with depression may be missed. Even more troubling, a wrong decision might harm people. While statistical methods will not eliminate wrong decisions, understanding statistical methods will allow you to reduce the number of wrong decisions you make. You are taking this course because the professionals in your discipline recognize that statistical methods improve decision making and make us better at our professions.

Reading Question

10. Why do many disciplines require students to take a statistics course? Taking a statistics course

a. is a way to employ statistics instructors, which is good for the economy.

b. can help people make better decisions in their chosen professions.

STATISTICS AND THE HELPING PROFESSIONS

When suffering from a physical or mental illness, we expect health professionals (e.g., medical doctors, nurses, clinical psychologists, and counselors) to accurately diagnose us and then prescribe effective treatments. We expect them to ask us detailed questions and then to use our answers (i.e., the data) to formulate a diagnosis. Decades of research has consistently found that health professionals who use statistics to make their diagnoses are more accurate than those who rely on their personal experience or intuition (e.g., Grove & Meehl, 1996).

For example, lawyers frequently ask forensic psychologists to determine if someone is likely to be violent in the future. In this situation, forensic psychologists typically review the person's medical and

criminal records as well as interview the person. Based on the records and the information gained during the interview, forensic psychologists make a final judgment about the person's potential for violence in the future. While making their professional judgment, forensic psychologists weigh the relative importance of the information in the records (i.e., the person's behavioral history) and the information obtained via the interview. This is an extremely difficult task. Fortunately, through the use of statistics, clinicians have developed methods that enable them to optimally gather and interpret data. One concrete example is the Violence Risk Appraisal Guide (Harris, Rice, & Quinsey, 1993). The guide is a list of questions that the psychologist answers after reviewing someone's behavioral history and conducting an interview. The answers to the guide questions are mathematically combined to yield a value that predicts the likelihood of future violence. Research indicates that clinicians who use statistical approaches such as the Violence Risk Appraisal Guide make more accurate clinical judgments than those who rely solely on their own judgment (Yang, Wong, & Coid, 2010). Today, statistical procedures help psychologists predict many things, including violent behavior, academic success, marital satisfaction, and work productivity. In addition to enabling us to make better predictions, statistical procedures also help professionals determine which medical or behavioral treatments are most effective.

Reading Question

11. Decades of research indicates that professionals in the helping professions make better decisions when they rely on
   a. statistics.
   b. their intuition and clinical experience.

HYPOTHESIS TESTING, EFFECT SIZE, AND CONFIDENCE INTERVALS

The statistical decisions you will make in this course revolve around specific hypotheses. A primary purpose of this book is to introduce the statistical process of null hypothesis significance testing (NHST), *a formal multiple-step procedure for evaluating the likelihood of a prediction, called a null hypothesis.* Knowledge of null hypothesis significance testing, also called significance testing or hypothesis testing, is fundamental to those working in the behavioral sciences, medicine, and the counseling professions. In later chapters, you will learn a variety of statistics that test different hypotheses. All the hypothesis testing procedures that you will learn are needed because of one fundamental problem that plagues all researchers—namely, the problem of sampling error. For example, researchers evaluating a new depression treatment want to know if it effectively lowers depression in all people with depression, called the population of people with depression. However, researchers cannot possibly study every depressed person in the world. Instead, researchers have to study a subset of this population, perhaps a sample of 100 people with depression. *The purpose of any sample is to represent the population from which it came.* In other words, if the 100 people with depression are a good sample, they will be similar to the population of people with depression. Thus, if the average score on a clinical assessment of depression in the population is 50, the average score of a good sample will also be 50. Likewise, if the ratio of women with depression to men with depression is 2:1 in the population, it will also be 2:1 in a good sample. Of course, you do not really expect a sample to be exactly like the population. *The differences between a sample and the population create* sampling error.

Reading Question

12. All hypothesis testing procedures were created so that researchers could
   a. study entire populations rather than samples.
   b. deal with sampling error.

Reading Question

13. If a sample represents a population well, it will
   (a.) respond in a way that is similar to how the entire population would respond.
   b. generate a large amount of sampling error.

While null hypothesis significance testing is extremely useful, it has limitations. Therefore, another primary purpose of this book is to describe these limitations and how researchers address them by using two additional statistical procedures. Effect sizes *describe the magnitude of a study's results,* helping researchers determine if a research result is large enough to be useful or if it is too small to be meaningful in "real-world" situations. Confidence intervals *identify the wide range of plausible values that might occur if sample results are applied to the entire population.* Each of these statistical procedures helps researchers give meaning to the results of a significance test. In fact, the American Psychological Association (APA) publication manual recommends that researchers use effect sizes and confidence intervals whenever significance tests are used (American Psychological Association, 2010). These three statistical procedures are most beneficial when they are used side by side.

Reading Question

14. Effect sizes and confidence intervals help researchers
   a. interpret (i.e., give meaning to) the results of significance tests.
   b. address the limitations of significance tests.
   (c.) do both of the above.

TESTING CAUSAL HYPOTHESES

While this book's main goal is teaching how to use the statistical procedures of hypothesis testing, effect sizes, and confidence intervals, you should know that there is a lot more to *causal* hypothesis testing than the statistics covered in this text. In many research situations, scientists want to know if manipulating one variable (the independent variable, or IV) *causes* a change in a second variable (the dependent variable, or DV). Testing *causal* hypotheses is particularly difficult because it requires carefully designed experiments. In these experiments, researchers must (1) manipulate the IV, (2) measure the DV after IV manipulation, (3) control for extraneous variables, and (4) provide evidence of a "significant" relationship between the IV manipulation and the DV score. For example, if we wanted to test the causal hypothesis that cell phone use while driving causes poorer driving performance, we would need to manipulate the IV (i.e., cell phone use) by having people operate a driving simulator while talking on a cell phone and also while not using a cell phone. Then, we would need to measure the DV of driving performance (e.g., braking reaction time or number of times people swerve out of their lane) when using a cell versus not. In order for us to feel confident that using the cell phone caused poorer driving performance, we would need to know that the two groups of people were equally good drivers and driving in equally challenging driving conditions in terms of traffic density, weather, destination, and so on. In other words, we need to make sure the test is "fair" in that the only difference between the two groups of drivers is whether or not they were using a cell phone while they were driving. Finally, only after carefully manipulating the IV, measuring the DV, and controlling extraneous variables do we use statistics to determine if the driving performances of those using cell phones versus not are so different that it justifies concluding that cell phone use while driving *causes* poorer driving performance. While the statistics you will learn in this text are a necessary component of testing causal hypotheses, they are not all you need to know. Causal hypothesis testing also requires mastery of experimental design. In a research methods course, you will learn how to design "fair" experiments that enable you to use the statistical procedures taught in this text to test causal hypotheses.

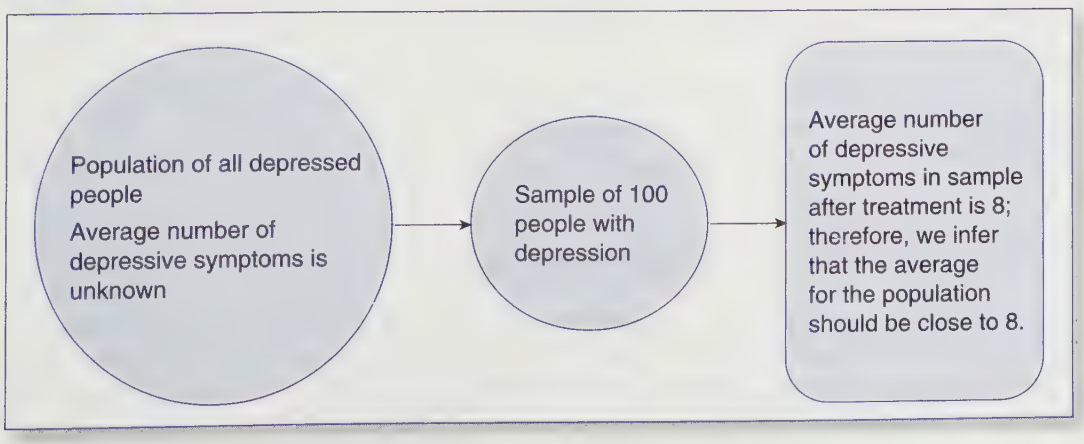
Reading Question

15. Testing casual hypotheses requires knowing how to
    a. use statistics.
    b. use research methods to design "fair" experiments.
    c. both of the above.

POPULATIONS AND SAMPLES

Suppose that a researcher studying depression gave a new treatment to a sample of 100 people with depression. Figure 1.1 is a pictorial representation of this research scenario. The large circle on the left represents a **population**, *a group of all things that share a set of characteristics*. In this case, the "things" are people, and the characteristic they all share is depression. Researchers want to know what the mean depression score for the population would be if all people with depression were treated with the new depression treatment. In other words, researchers want to know the **population parameter**, *the value that would be obtained if the entire population were actually studied*. Of course, the researchers don't have the resources to study every person with depression in the world, so they must instead study a **sample**, *a subset of the population that is intended to represent the population*. In most cases, the best way to get a sample that accurately represents the population is by taking a **random sample** from the population. When taking a **random sample**, *each individual in the population has the same chance of being selected for the sample*. In other words, while researchers want to know a population parameter, their investigations usually produce a **sample statistic**, *the value obtained from the sample*. The researchers then use the sample statistic value as an estimate of the population parameter value. The researchers are making an *inference* that the sample statistic is a value similar to the population parameter value based on the premise that the characteristics of those in the sample are similar to the characteristics of those in the entire population. *When researchers use a sample statistic to infer the value of a population parameter,* it is called **inferential statistics.** For example, a researcher studying depression wants to know how many depressive symptoms are exhibited by people in the general population. He can't survey everyone in the population, and so he selects a random sample of people

Figure 1.1 A Pictorial Representation of Using a Sample to Estimate a Population Parameter (i.e., Inferential Statistics)



from the population and finds that the average number of symptoms in the sample is 8 (see Figure 1.1). If he then inferred that the entire population of people would have an average of 8 depressive symptoms, he would be basing his conclusion on inferential statistics. It should be clear to you that if the sample did not represent the population well (i.e., if there was a lot of sampling error), the sample statistic would NOT be similar to the population parameter. In fact, **sampling error** is defined as *the difference between a sample statistic value and an actual population parameter value.*

Reading Question

16. The value obtained from a population is called a
    a. statistic.
    b. parameter.

Reading Question

17. Parameters are
    a. always exactly equal to sample statistics.
    b. often estimated or inferred from sample statistics.

Reading Question

18. When a statistic and parameter differ,
    a. it is called an inferential statistic.
    b. there is sampling error.

The researchers studying depression were using inferential statistics because they were using data from a sample to infer the value of a population parameter. The component of the process that makes it inferential is that researchers are using data they actually have to estimate (or infer) the value of data they don't actually have. In contrast, researchers use **descriptive statistics** *when their intent is to describe the data that they actually collected.* For example, if a clinical psychologist conducted a study in which she gave some of her clients a new depression treatment and she wanted to describe the average depression score of only those clients who got the treatment, she would be using descriptive statistics. Her intent is only to describe the results she observed in the clients who actually got the treatment. However, if she then wanted to estimate what the results would be if she were to give the same treatment to additional clients, she would then be performing inferential statistics.

Reading Question

19. Researchers are using descriptive statistics if they are using their results to
    a. estimate a population parameter.
    b. describe the data they actually collected.

Reading Question

20. Researchers are using inferential statistics if they are using their results to
    a. estimate a population parameter.
    b. describe the data they actually collected.

INDEPENDENT AND DEPENDENT VARIABLES

Researchers design experiments to test if one or more variables cause changes to another variable. For example, if a researcher thinks a new treatment reduces depressive symptoms, he could design an experiment to test this prediction. He might give a sample of people with depression the new treatment and withhold the treatment from another sample of people with depression. Later, if those who received the new treatment had lower levels

of depression, he would have evidence that the new treatment reduces depression. In this experiment, the type of treatment each person received (i.e., new treatment vs. no treatment) is the **independent variable (IV)**. In this study, the experimenter manipulated the IV by giving one sample of people with depression the new treatment and another sample of people with depression a placebo treatment that is not expected to reduce depression. In this experiment, the IV has two **IV levels:** (1) the new treatment and (2) the placebo treatment. The main point of the study is to determine if the two different IV levels were differentially effective at reducing depressive symptoms. More generally, *the IV is a variable with two or more levels that are expected to have different impacts on another variable.* In this study, after both samples of people with depression were given their respective treatment levels, the amount of depression in each sample was compared by counting the number of depressive symptoms in each person. In this experiment, the number of depressive symptoms observed in each person is the **dependent variable (DV)**. Given that the researcher expects the new treatment to work and the placebo treatment not to work, he expects the new treatment DV scores to be lower than the placebo treatment DV scores. More generally, *the DV is the outcome variable that is used to compare the effects of the different IV levels.*

Reading Question

21. The IV (independent variable) in a study is the
   a. variable expected to change the outcome variable.
   b. outcome variable.

Reading Question

22. The DV (dependent variable) in a study is the
   a. variable expected to change the outcome variable.
   b. outcome variable.

In true experiments, those in which researchers manipulate a variable so that some participants have one value and others have a different value, the manipulated variable is always referred to as the IV. For example, if a researcher gives some participants a drug (Treatment A) and others a placebo (Treatment B), this manipulation defines the IV of treatment as having two levels—namely, drug and placebo. However, in this text, we also use the IV in a more general way. The IV is any variable predicted to influence another variable even when the IV was not manipulated. For example, if a researcher predicted that women would be more depressed than men, we will refer to gender as the IV because it is the variable that is expected to influence the DV (i.e., depression score). If you take a research methods course, you will learn an important distinction between manipulated IVs (e.g., type of treatment: drug vs. placebo) and *measured* IVs (e.g., gender: male vs. female). Very briefly, the ultimate goal of science is to discover causal relationships, and manipulated IVs allow researchers to draw causal conclusions while measured IVs do not. You can learn more about this important distinction and its implications for drawing causal conclusions in a research methods course.

Reading Question

23. All studies allow you to determine if the IV causes changes in the DV.
   a. True
   b. False

? words

SCALES OF MEASUREMENT

All research is based on measurement. For example, if researchers are studying depression, they will need to devise a way to measure depression accurately and reliably. The way a variable is measured has a direct impact on the types of statistical procedures that can be used to analyze that variable. Generally speaking, researchers want to devise measurement procedures that are as precise as possible

because more precise measurements enable more sophisticated statistical procedures. Researchers recognize four different **scales of measurement** that vary in their degree of measurement precision: (1) nominal, (2) ordinal, (3) interval, and (4) ratio (Stevens, 1946). Each of these scales of measurement is increasingly more precise than its predecessor, and therefore, each succeeding scale of measurement allows more sophisticated statistical analyses than its predecessor.

Reading Question

24. The way a variable is measured
    a. determines the kinds of statistical procedures that can be used on that variable.
    b. has very little impact on how researchers conduct their statistical analyses.

For example, researchers could describe depression using a nominal scale by categorizing people with different kinds of major depressive disorders into groups, including those with melancholic depression, atypical depression, catatonic depression, seasonal affective disorder, or postpartum depression. **Nominal scales** of measurement *categorize things into groups that are qualitatively different from other groups.* Because nominal scales of measurement involve categorizing individuals into qualitatively distinct categories, they yield **qualitative** data. In this case, clinical researchers would interview each person and then decide which type of major depressive disorder each person has. With nominal scales of measurement, it is important to note that the categories are not in any particular order. A diagnosis of melancholic depression is not considered "more depressed" than a diagnosis of atypical depression. With all other scales of measurement, the categories are ordered. For example, researchers could also measure depression on an ordinal scale by ranking individual people in terms of the severity of their depression. **Ordinal scales** of measurement also categorize people into different groups, but on ordinal scales, these groups are rank ordered. In this case, researchers might interview people and diagnose them with a "mild depressive disorder," "moderate depressive disorder," or "severe depressive disorder." An ordinal scale clearly indicates that people *differ in the amount of something they possess.* Thus, someone who was diagnosed with mild depressive disorder would be less depressed than someone diagnosed with moderate depressive disorder. Although ordinal scales rank diagnoses by severity, they do not quantify how much more depressed a moderately depressed person is relative to a mildly depressed person. To make statements about how much more depressed one person is than another, an interval or ratio measurement scale is required. Researchers could measure depression on an interval scale by having people complete a multiple-choice questionnaire that is designed to yield a score reflecting the amount of depression each person has. **Interval scales** of measurement *quantify how much of something people have.* While the ordinal scale indicates that some people have more or less of something than others, the interval scale is more precise indicating exactly *how much* of something someone has. Another way to think about this is that for interval scales, the intervals between categories are equivalent, whereas for ordinal scales, the intervals are not equivalent. For example, on an ordinal scale, the interval (or distance) between a mild depressive disorder and a moderate depressive disorder may not be the same as the interval between a moderate depressive disorder and a severe depressive disorder. However, on an interval scale, the distances between values are equivalent. If people completed a well-designed survey instrument that yielded a score between 1 and 50, the difference in the amount of depression between scores 21 and 22 would be the same as the difference in the amount of depression between scores 41 and 42. Most questionnaires used for research purposes yield scores that are measured on an interval scale of measurement. **Ratio scales** of measurement also *involve quantifying how much of something people have, but a score of zero on a ratio scale indicates that the person has none of the thing being measured.* For example, if people are asked how much money they earned last year, the income variable would be measured on a ratio scale because not only are the intervals between values equivalent, but there also is an absolute zero point. A value of zero means the complete absence of income last year. Because they involve quantifying how much of something an individual has, interval and ratio scales yield **quantitative** data. Interval and ratio scales are similar in that they both determine how much of something

someone has but some interval scales can yield a negative number, while the lowest score possible on a ratio scale is zero. Within the behavioral sciences, the distinction between interval and ratio scales of measurement is not usually very important. Researchers typically use the same statistical procedures to analyze variables measured on interval and ratio scales of measurement.

Although most variables can be easily classified as nominal, ordinal, or interval/ratio, some data are more difficult to classify. Researchers often obtain data by asking participants to answer questions on a survey. These survey responses are then combined into a single measure of the construct. For example, participants may answer a series of questions related to depression, and then the researcher would combine those questions into a single depression score. Although there is not complete agreement among statisticians, most researchers classify summed scores from questionnaires and surveys as interval data (e.g., Carifio & Perla, 2007). Thus, in this course, summed scores from surveys will be considered interval/ratio data.

Reading Question

25. Researchers typically treat summed questionnaire/survey scores as which scale of measurement?

a. Nominal scale of measurement

b. Ordinal scale of measurement

c. Interval scale of measurement

When trying to identify the scale of measurement of a variable, it can also be helpful to think about what each scale of measurement allows you to do. For example, if you can only count the number of things in a given category, you know that you have a nominal scale. Table 1.1 summarizes what you can do with each type of scale and provides examples of each scale of measurement.

Table 1.1 The Four Scales of Measurement, What They Allow, and Examples

Scale of Measurement	*What the Scale Allows You to Do*	*Examples*
Nominal	COUNT the number of things within different categories	*Pets:* 5 dogs, 12 cats, 7 fish, 2 hamsters
		Marital status: 12 married, 10 divorced, 2 separated
Ordinal	COUNT & RANK some things as having more of something than others (but NOT QUANTIFY how much of it they have)	*Annual income:* above average, average, or below average
		Speed (measured by place of finish in a race): 1st, 2nd, 3rd, etc.
Interval	COUNT, RANK, & QUANTIFY how much of something there is, but a score of zero does not mean the absence of the thing being measured	*Temperature:* –2°F, 98°F, 57°F; 0°F is not the absence of heat
Ratio	COUNT, RANK, & QUANTIFY how much of something there is, and a score of zero means the absence of the thing being measured	*Annual income:* $25,048, $48,802, $157,435, etc.
		Number of text messages sent in a day: 0, 3,351, 15, etc.

Reading Question

26. The scale of measurement that quantifies the thing being measured (i.e., indicates *how much* of it there is) is _____ scale(s) of measurement.
    a. the nominal
    b. the ordinal
    c. both the interval and ratio

Reading Question

27. The scale of measurement that categorizes objects into different kinds of things is _____ scale(s) of measurement.
    a. the nominal
    b. the ordinal
    c. both the interval and ratio

Reading Question

28. The scale of measurement that indicates that some objects have more of something than other objects but not how much more is _____ scale(s) of measurement
    a. the nominal
    b. the ordinal
    c. both the interval and ratio

DISCRETE VERSUS CONTINUOUS VARIABLES

Variables can also be categorized as discrete or continuous. A **discrete variable** *only occurs in whole units rather than fractions of units*. For example, the variable "number of siblings" is a discrete variable because someone can only have a whole number of siblings (e.g., no one can have 2.7 siblings). A **continuous variable** *occurs in fractions of units*. For example, the variable "time to complete a test" is a continuous variable because someone can take a fraction of minutes to complete a test (e.g., 27.39 minutes). Nominal and ordinal variables are always discrete variables. Interval and ratio variables can be either discrete or continuous.

Reading Question

29. If a variable can be measured in fractions of units, it is a _____ variable.
    a. discrete
    b. continuous ...

GRAPHING DATA

Graphing often helps you understand your data. For example, if you were looking at the number of siblings college students have, you could begin by looking at a graph to determine how many siblings most students have. Inspection of the graph also allows you to find out if there is anything odd in the data file that requires further examination. For example, if you graphed the data and found that most people reported having between 0 and 4 siblings but one person reported having 20 siblings, you should probably investigate to determine if that 20 was an error.

There are three basic types of graphs that we use for most data: (1) **bar graphs**, (2) **histograms**, and (3) **line graphs**. The names of the first two are a bit misleading because both are created using bars. The only difference between a bar graph and a histogram is that in a bar graph, the bars do not touch while the bars do touch in a histogram. In general, use bar graphs when the data are discrete or qualitative. The space between the bars of a bar graph emphasize that there are no possible values between any two categories (i.e., bars). For example, when graphing the number of children in a family, a bar graph is appropriate because there is no possible value between any two categories (e.g., you cannot have 1.5 children). When the data are continuous, use a histogram. For example, if you are graphing the variable "time to complete a test" and creating a bar for each minute category, the bars would touch to indicate that the variable we are graphing is continuous (i.e., 27.46 minutes is possible).

Reading Question

30. What type of graph is used for discrete data or qualitative data?
   a. Bar graph
   b. Histogram

Reading Question

31. What type of graph is used for continuous data?
   a. Bar graph
   b. Histogram

Reading Question

32. In bar graphs, the bars _____.
   a. touch
   b. don't touch

Reading Question

33. In histograms, the bars _____.
   a. touch
   b. don't touch

To create either a bar graph or a histogram, you should put categories on the x-axis and the number of scores in a particular category (i.e., the frequency) on the y-axis. For example, suppose we asked 19 students how many siblings they have and obtained the following responses:

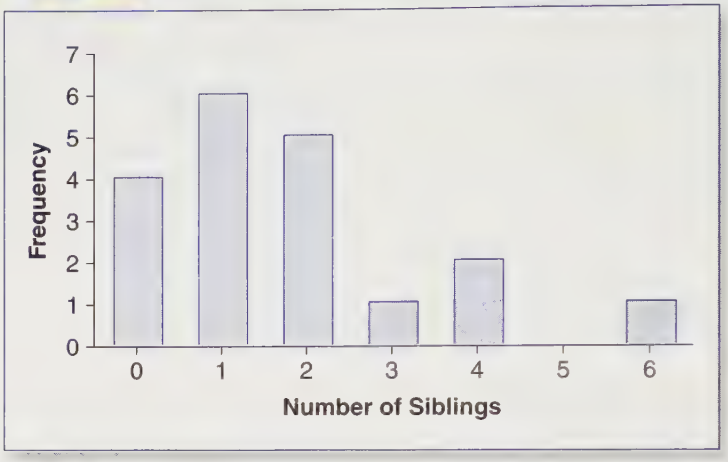
0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 6

To graph these responses, you would list the range of responses to the question, "How many siblings do you have?" on the x-axis (i.e., in this case, 0 through 6). The y-axis is the frequency within each category. For each response category, you will draw a bar with a height equal to the number of times that response was given. For example, in the bar graph (Figure 1.2), 4 people said they had 0 siblings, and so the bar above the 0 has a height of 4.

Reading Question

34. Use the graph to determine how many people said they had 1 sibling.
   a. 4
   b. 5
   c. 6

Figure 1.2 Bar Graph of Variable, Number of Siblings, Collected From a Sample of 19 Students

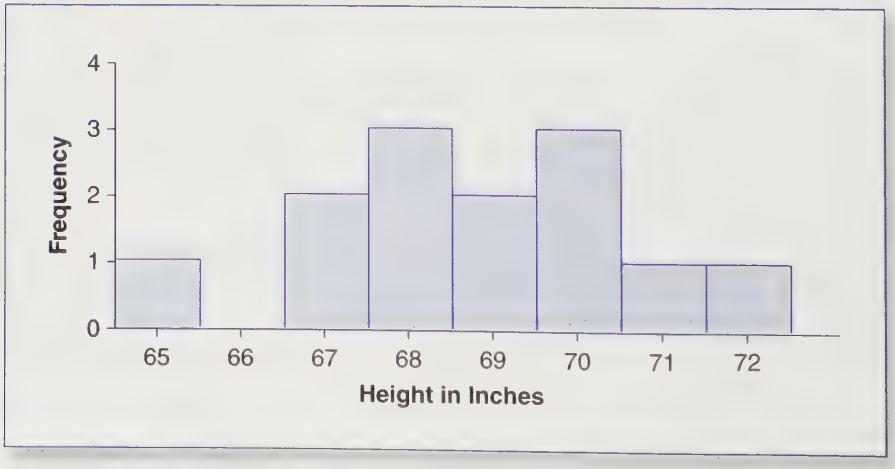


The procedure for creating a histogram is similar to that for creating a bar graph. The only difference is that the bars should touch. For example, suppose that you recorded the height of players on a volleyball team and obtained the following heights rounded to the nearest inch:

65, 67, 67, 68, 68, 68, 69, 69, 70, 70, 70, 71, 72

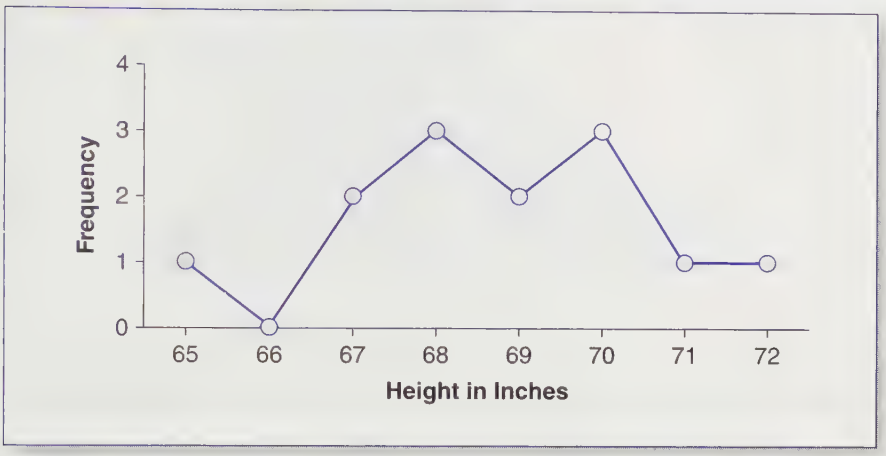
Height in inches is continuous because there are an infinite number of possible values between any two categories (e.g., between 68 and 69 inches). The data are continuous so we create a histogram (i.e., we allow the bars to touch) (Figure 1.3).

Figure 1.3 Frequency Histogram of Variable, Height in Inches, Collected From a Sample of 13 Volleyball Players



Whenever a histogram is appropriate, you may also use a **line graph** in its place. To create a line graph, you use dots to indicate frequencies and connect adjacent dots with lines (Figure 1.4).

Figure 1.4 Frequency Line Graph of Variable, Height in Inches, Collected From a Sample of 13 Volleyball Players



Whether the data are discrete or continuous should determine how the data are graphed. You should use a bar graph for discrete data and a histogram or a line graph for continuous data. Nominal data should be graphed with a bar graph. Throughout the text, we will use these guidelines, but you should be aware of the fact that histograms and bar graphs are often used interchangeably outside of statistics classes.

Reading Question

35. Line graphs can be used whenever a _____ is appropriate.
   a. histogram
   b. bar graph

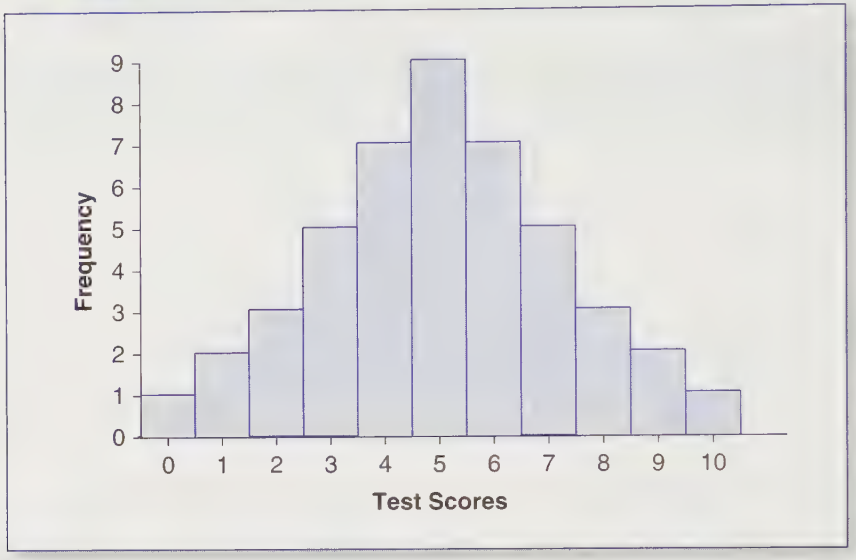
Reading Question

36. What type of graph should be used if the data are measured on a nominal scale?
   a. Histogram
   b. Bar graph

SHAPES OF DISTRIBUTIONS

A **distribution** is *a group of scores*. If a distribution is graphed, the resulting bar graph or histogram can have any "shape," but certain shapes occur so frequently that they have specific names. The most common shape you will see is a bell curve. The curve in Figure 1.5 resembles a bell-shaped distribution. Bell-shaped distributions are also called *normal distributions* or *Gaussian distributions*.

Figure 1.5 Frequency Histogram of Test Scores That Form a Normal Curve



One important characteristic of normal distributions is that most of the scores pile up in the middle, and as you move further from the middle, the frequency of the scores gets less. In addition, normal distributions are symmetrical in that the right and left sides of the graph are identical.

For the purposes of this book, you do not need to know the exact mathematical properties that define the normal curve. However, you should know that a normal curve looks bell shaped and symmetrical. You will use the normal curve frequently in this book.

The normal curve is important because many variables, when graphed, have a normal shape, and this fact will be very important in later chapters. While normal curves are common, there are specific ways for graphs to deviate from a normal bell shape. Some of these deviations have specific names. For example, graphs can deviate from the bell shape because of **skew.** A skewed distribution is asymmetrical, meaning the right and left sides are not identical. Instead, the scores are shifted such that most of them occur on one side of the peak with fewer scores on the other side of the scale. For example, the distributions in Figures 1.6 and 1.7 are both skewed, but in different ways. The positively skewed distribution (Figure 1.6) has the majority of the scores on the low end of the distribution with fewer scores on the higher end. The negatively skewed distribution is the opposite. Distinguishing between positive and negative skew is as easy as noticing which side of the distribution has the longer "tail" (i.e., which side takes longer to descend from the peak to zero frequency). In positively skewed distributions, the longer tail points toward the right, or the positive side of the x-axis. In negatively skewed distributions (Figure 1.7), the longer tail points toward the left, or the negative side of the x-axis. There are statistics that you can compute to quantify exactly how skewed a distribution is (see Field, 2013, for an excellent discussion), but we will just eyeball the graphs to determine if they deviate from normal.

Figure 1.6 Positively Skewed Distribution

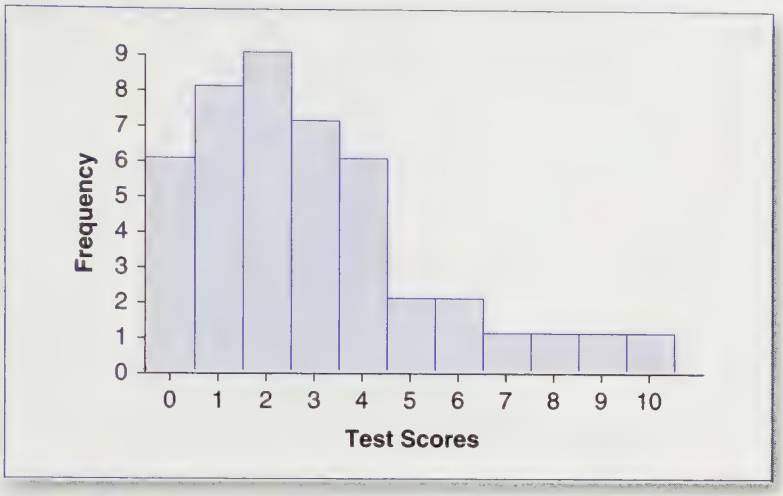
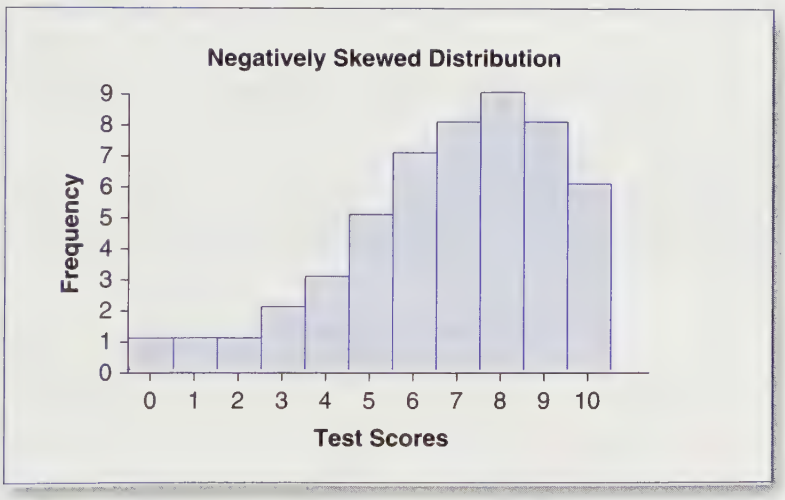


Figure 1.7 Negatively Skewed Distribution of Scores



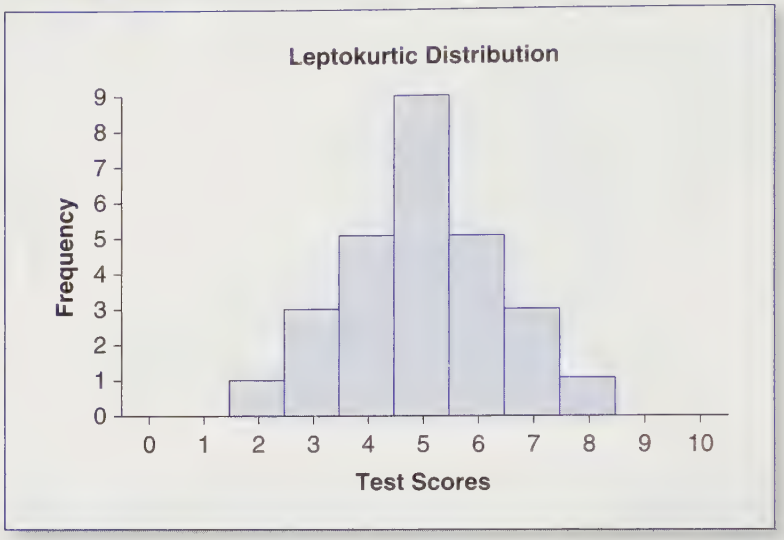
Reading Question

37. The scores on an exam are distributed such that most scores are low (between 30% and 50%), but a couple of people had very high scores (i.e., above 95%). How is this distribution skewed?

(a.) Positively skewed

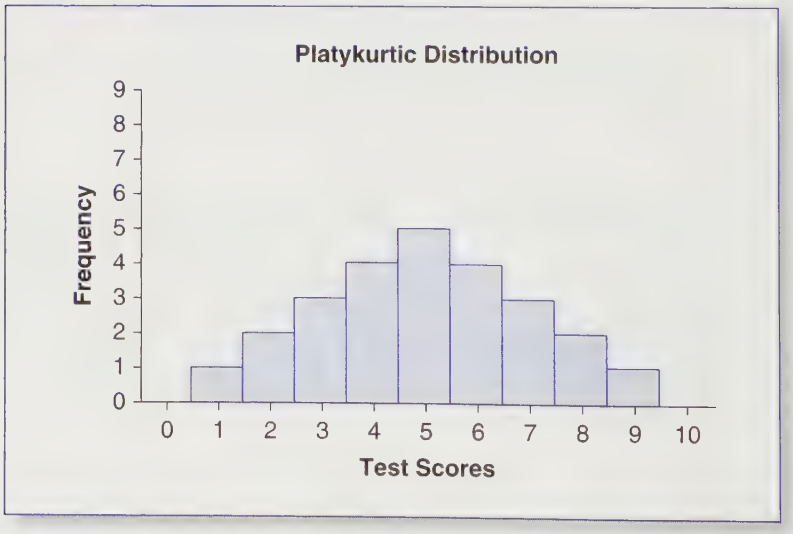
b. Negatively skewed

Figure 1.8 Example of a Leptokurtic Distribution



Distributions also vary in kurtosis, which is the extent to which they have an exaggerated peak versus a flatter appearance. Distributions that have a higher, more exaggerated peak than a normal curve are called leptokurtic, while those that have a flatter peak are called platykurtic. Figures 1.8 and 1.9 display a

Figure 1.9 Example of a Platykurtic Distribution



leptokurtic and platykurtic distribution, respectively. As with skew, there are ways to quantify kurtosis in a distribution (again, see Field, 2013), but we will just eyeball it in this book.

Reading Question

38. Distributions that are flatter than a normal distribution are called
   a. platykurtic.
   b. leptokurtic.

plat = flat

FREQUENCY DISTRIBUTION TABLES

Graphing data is typically the best way to see patterns in the data (e.g., normal, leptokurtic, or platykurtic). However, some precision is often lost with graphs. Therefore, it is sometimes useful to look at the raw data in a **frequency distribution table**. To create a frequency distribution table, you need to know the measurement categories as well as the number of responses within a given measurement category. For example, suppose that a market researcher asked cell phone users to respond to the following statement: "I am very happy with my cell phone service provider." People were asked to respond with 1 = *strongly agree*, 2 = *agree*, 3 = *neither agree nor disagree*, 4 = *disagree*, or 5 = *strongly disagree*. The responses are listed below:

1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5

It is probably obvious that a string of numbers like the one earlier is not a particularly useful way to present data. A frequency distribution table organizes the data, so it is easier to interpret; one is shown in Table 1.2.

The first column (*X*) represents the possible response categories. People *could* respond with any number between 1 and 5; therefore, the *X* column (i.e., the measurement categories) must include all of the *possible* response values—namely, 1 through 5. In this case, we chose to put the categories in ascending order from 1 to 5, but they could also be listed in descending order from 5 to 1.

The next column (*f*) is where you record the frequency of each response. For example, 4 people gave responses of 5 (*strongly disagree*) and so a 4 is written in the "*f*" column across from the response category of 5 (*strongly disagree*).

Table 1.2 Frequency Distribution Table of the Variable "I Am Very Happy With My Cell Phone Service Provider"

	X	*f*
Strongly agree	1	2
Agree	2	4
Neither agree nor disagree	3	7
Disagree	4	6
Strongly disagree	5	4

Reading Question

39. The value for "*f*" represents the
   a. number of measurement categories.
   b. number of responses within a given measurement category.

Reading Question

40. In the above frequency table, how many people responded with an answer of 3?
   a. 2
   b. 4
   c. 7

SPSS

We will be using a statistical package called **SPSS (Statistical Package for the Social Sciences)** to conduct many of the statistical analyses in this course. Our instructions and screenshots were developed with Version 22. There are some minor differences between Version 22 and other versions, but you should have no difficulty using our instructions with other SPSS versions.

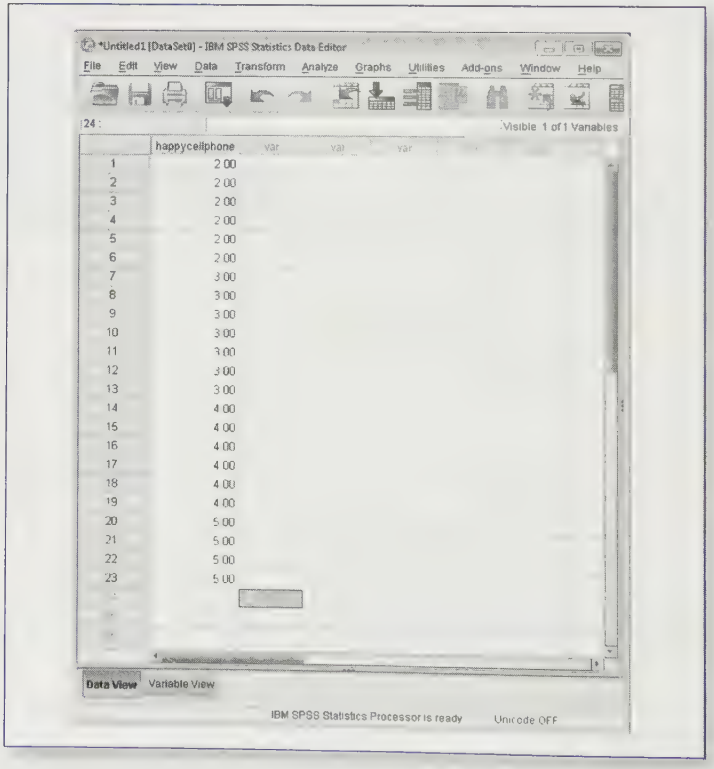
It is likely that your school has a site license for SPSS allowing you to access it on campus. Depending on your school's site license, you may also be able to access the program off campus. You may also purchase or "lease" a student or graduate version of SPSS for this course. Your instructor will tell you about the options available to you.

Data File

After you open SPSS, click on the Data View tab near the bottom left of the screen. Enter the data you want to analyze in a single column.

We have used the cell phone data from the previous page to help illustrate how to use SPSS. In Figure 1.10, a variable named "happycellphone" is shown at the top of the column of data. To add this variable name, double click on the blue box at the top of a column in the Data View screen.

Figure 1.10 Screenshot of SPSS Data Entry Screen



Doing so will take you to the Variable View screen. You can also access the Variable View screen by pressing the Variable View tab at the bottom left of the screen. In the first column and first row of the Variable View screen, type the name of the variable you want to appear in the data spreadsheet (e.g., happycellphone—the variable name cannot have spaces or start with a number). To go back to the Data View, click on the blue Data View tab at the bottom left of the screen.

The data file you created should look like the screenshot in Figure 1.10. The exact order of the data values is not important, but all 23 scores should be in a single column. As a general rule, all the data for a variable must be entered in a single column.

Reading Question

41. The Variable View screen is where you
   a. enter the variable names.
   b. enter the data.

Reading Question

42. The Data View screen is where you
   a. enter the variable names.
   b. enter the data.

Creating Frequency Distribution Tables and Graphs

SPSS can create frequency tables and graphs. To create a frequency graph of the data you just entered, do the following:

- From the Data View screen, click on Analyze, Descriptive Statistics, and then Frequencies.
- To create a graph, click on the Charts button and then choose the type of graph you want to create (Bar chart, Pie chart, or Histogram). Click on the Continue button.
- Be sure that the Display Frequency Tables box is checked if you want to create a frequency distribution table.
- Click on the OK button to create the frequency distribution table and graph.

After performing the steps outlined above, a frequency distribution graph and table will appear in the SPSS output screen. Use the SPSS output provided in Figure 1.11 to answer the following three questions.

Reading Question

43. How many people responded with a 3 to the question, "I am very happy with my cell phone provider?"
   a. 2
   b. 4
   c. 7

Reading Question

44. What percentage of the respondents answered the question with a response of 4?
   a. 30.4
   b. 26.1
   c. 17.4

round?

Figure 1.11 Annotated SPSS Frequency Table Output

Valid:
Each measurement category (X) in the data file

Frequency:
Number of responses in each measurement category

happycellphone

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	2	8.7	8.7	8.7
	2.00	4	17.4	17.4	26.1
	3.00	7	30.4	30.4	56.5
	4.00	6	26.1	26.1	82.6
	5.00	4	17.4	17.4	100.0
	Total	23	100.0	100.0	

Percent:
Percentage of responses in each measurement category: $(f/N)(100)$, where N = total number of participants in the data file

Valid Percent:
Percentage of responses in each measurement category: $(f/N)(100)$, where N = total number of participants who responded

Cumulative Percent:
Percentage of responses at or below a given measurement category

Reading Question

45. What percentage of the respondents answered the question with a response of 4 or a lower value?

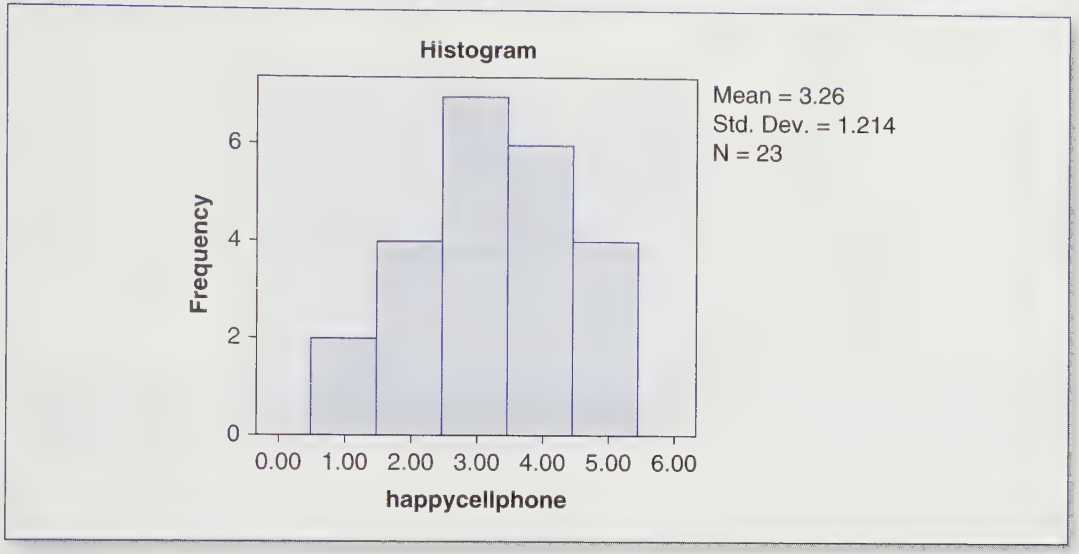
a. 56.5

b. 82.6

c. 100

Use the histogram in Figure 1.12 to answer the following two questions.

Figure 1.12 Frequency Histogram of "I Am Very Happy With My Cell Phone Service Provider" Data



Reading Question

46. What is the most common response in the data?

a. 2
b. 3
c. 4
d. 5

Reading Question

47. How many people responded with the most common response?

a. 7
b. 6
c. 5
d. 4

SPSS is a great tool for creating graphs to help you gain a better understanding of your data. However, it is not necessarily intended for creating presentation-quality graphs. You can customize graphs in SPSS by double clicking on the graph once you create it and then, once the image is open, double click on any aspect of the graph to change it. This is trickier than it sounds because there are a lot of options. We are not going to work on editing graphs in this book, but if you would like to edit graphs, you can use the help menu in SPSS to obtain further information. There are several other ways to create more advanced graphs in SPSS. You can explore these options by clicking on "Graphs" menu.

Reading Question

48. It is possible to change the appearance of graphs created by SPSS.
   a. True
   b. False

OVERVIEW OF THE ACTIVITY

In Activity 1.1, you will practice using the concepts introduced in this chapter, including samples, descriptive statistics, inferential statistics, populations, parameters, and sampling error. You will create frequency distribution tables by hand and using SPSS. When interpreting these tables, you will also learn about percentiles and how they can be obtained from a frequency distribution table. You will also create graphs by hand and using SPSS and describe their skew and kurtosis using the correct terminology. Finally, you will read research scenarios and determine what scale of measurement best describes the variables in the study.

Activity 1.1: Frequency Distributions

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Use common statistical terms correctly in a statistical context
- Construct a frequency distribution table from a bar graph
- Interpret data from a frequency distribution
- Use SPSS to create a frequency table
- Sketch a frequency distribution
- Identify distributions that are bell shaped, positively skewed, negatively skewed, leptokurtic, and platykurtic
- Identify nominal, ordinal, and interval/ratio variables in research scenarios
- Identify discrete and continuous variables in research scenarios

THERAPEUTIC TOUCH

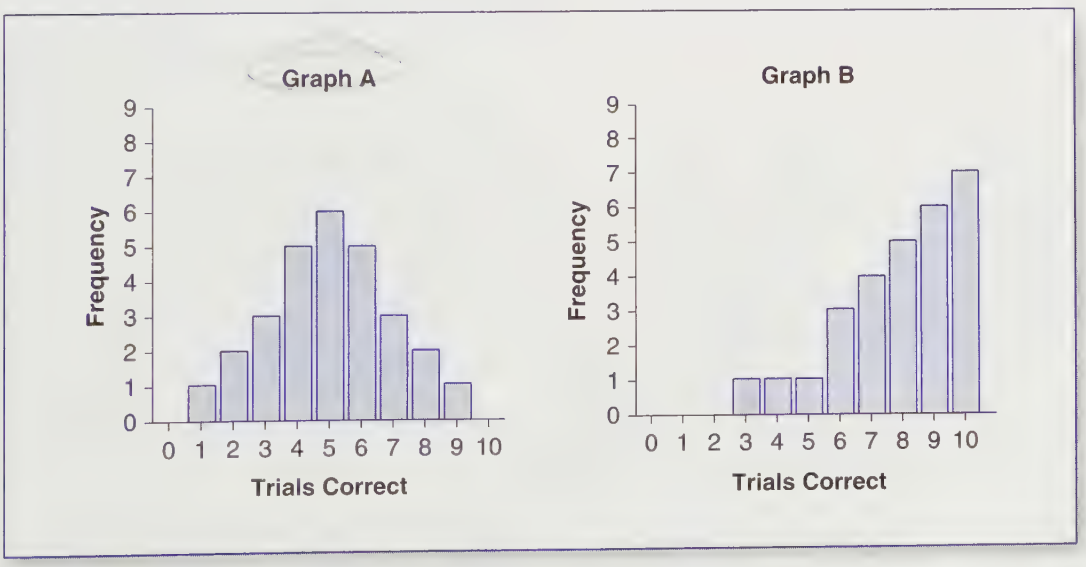
There is quite a bit of evidence that human touch is beneficial to our psychological and physical health. Hugs are associated with lower blood pressure, skin-to-skin contact can help preterm infants gain weight, and touch can improve immune system functioning (e.g., Field, 2010). Although there is little doubt of the benefits of physical touch, a treatment known as "therapeutic touch" (TT) is far more controversial. Therapeutic touch involves no actual physical contact. Instead, practitioners use their hands to move "human energy fields" (HEFs) in an attempt to promote healing. Proponents of this approach claim that it can help with relaxation, reduce pain, and improve the immune system.



Emily Rosa (who was just 9 years old at the time) and her colleagues (including her parents) investigated the basis of these TT claims by putting a sample of actual TT practitioners to the test (Rosa, Rosa, Sarner, & Barrett, 1998). In their study, Rosa and colleagues designed a method to determine if TT practitioners could actually detect HEFs. As the figure to the right illustrates, individual practitioners sat at a table facing a large divider that prevented them from seeing their own hands or Emily. The practitioners placed both of their hands through the divider on the table, palms up. Practitioners were told to indicate whether Emily was holding her hand above their right or left hand. Emily began each trial by flipping a coin to determine where to place her hand. She then placed her hand 8 to 10 cm above one of the practitioner's hands. The practitioners had to "sense" the HEF allegedly emanating from Emily's hand to determine if Emily's hand was over their right hand or left hand. Each practitioner went through a total of 10 of these trials.

If the TT practitioners can actually sense HEFs, they should be able to choose the correct hand far better than chance (i.e., 5 out of 10 times). However, if they really can't detect HEFs and the practitioners were really guessing, you would expect them to choose the correct hand *an average* of 5 out of 10 times. Some may get more than 5 correct and others may get less than 5 correct, but the most common number of correct answers would be about 5 of 10, *if the practitioners were guessing*.

1. Which of the following graphs represents the results you would expect *if the practitioners were guessing*? Explain your answer.



2. As mentioned previously, the researchers had a sample of TT practitioners participate in the experiment described earlier. They used the results from this sample to infer what the results would be if they had collected data from the entire population of TT practitioners. The purpose of their study was

   a. descriptive.

   b. inferential.

3. Use three of the following terms to fill in the blanks: parameters, statistics, inferential, descriptive, sampling error.

   If the sample of TT practitioners represented the population of TT practitioners well, the sample _____ would be similar to the population _____ and the study would have a relatively small amount of _____.

4. After the experiment was complete, the researchers counted the number of correct responses out of the 10 possible that were generated by each participant. The number of correct responses ranged between a low of 1 correct to a high of 8 correct. The variable "number of correct responses out of 10 trials" is measured on which scale of measurement?

   a. Nominal

   b. Ordinal

   c. Interval/ratio

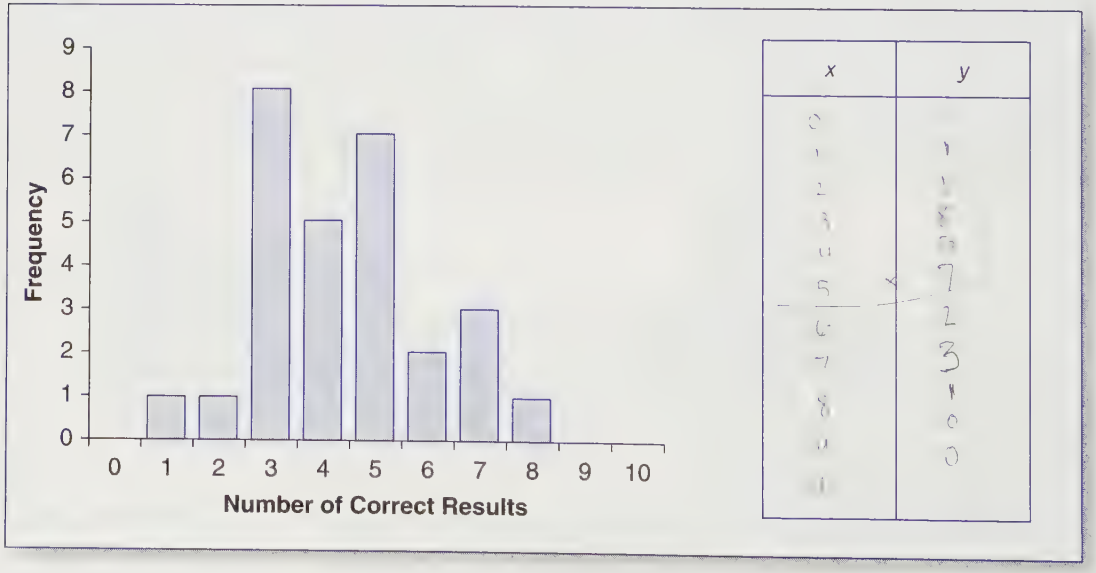
5. Is the number of correct responses out of 10 a continuous or a discrete variable?

   a. Continuous

   b. Discrete

The following bar graph is an accurate re-creation of the actual data from the experiment. The graph is a frequency distribution of the number of correct responses generated by each practitioner out of 10 trials. Use these data to answer the following questions:

6. Create a frequency distribution table based on the graph.



7. How many practitioners were in the sample?
   a. 8
   b. 10
   c. 28

8. How many practitioners did *better* than chance (i.e., did better than 5 correct out of 10)?
   a. 3
   b. 6
   c. 13

9. What *percentage* of the practitioners performed *at or below* chance?
   a. 100
   b. 78.6
   c. 53.6

10. Do the data support the conclusion that TT practitioners can detect HEFs or do the data support the conclusion that they cannot and instead are guessing?
    a. Yes, many of the practitioners performed above chance level. Although the other practitioners could not detect the HEFs, the people who scored above chance could detect HEFs.
    b. No, most practitioners performed at or below chance levels. This suggests that, generally, the TT practitioners were not able to detect the HEFs.

11. Some of the TT practitioners were correct on 6, 7, or 8 of the trials. What should the researchers do next?
    a. Conclude that these four individuals really can detect HEFs and encourage them to continue using TT to treat people.
    b. Do the study again with the same people and see if they can replicate their above-chance performance.

GENERAL SOCIAL SURVEY

Every 2 years, the National Opinion Research Center asks a random sample of adults in the United States to complete the **General Social Survey** (GSS). All of the GSS data are available at www.norc.org. You will be using a small portion of the GSS that we placed in a file titled "gss2010.sav." You can access this file on the textbook website (http://www.sagepub .com/carlson/study/resources.htm). Load this file into SPSS.

Part of the GSS assesses respondents' science knowledge. In 2010, respondents answered questions from a variety of different sciences, such as "True or False. Antibiotics kill viruses as well as bacteria" and "True or False. Lasers work by focusing sound waves." For this assignment, we created the variable "ScientificKnowledge" by summing the total number of correct answers each participant gave to 10 science questions. The resulting "ScientificKnowledge" variable was measured on a ratio scale and had a possible range of 0 to 10 correct answers.

Use SPSS to create a frequency distribution table and graph of "ScientificKnowledge" scores. To create a frequency distribution table and graph, do the following:

- From the Data View screen, click on Analyze, Descriptive Statistics, and then Frequencies.
- Move the variable(s) of interest into the Variable(s) box. In this case, you will move "ScientificKnowledge" into the Variable(s) box.

- Make sure the Display Frequency Tables box is checked.
- To create a graph, click on the Charts button and then choose Bar chart. Click on the Continue button.
- Click on the OK button to create the frequency distribution table and graph.

12. Use the frequency distribution table you created in SPSS to determine how many people responded to the "ScientificKnowledge" questions.

937

13. How many people got all 10 questions right?

18

14. What *percentage* of people got all 10 questions right? You could compute this percentage yourself, but SPSS has already computed it for you. You will notice that there are two percentage columns in SPSS. One is labeled "Percent" and the other is labeled "Valid Percent." These two columns differ in what was used as the denominator when computing the percentage. For the Percent column, the denominator is everyone in the data file, regardless of whether they answered the question. For the Valid Percent column, the denominator is the number of people who answered the question. For this course, we will always use the Valid Percent column.

5.1

15. How many people got all 10 questions wrong?

14

16. What *percentage* of people got all 10 questions wrong?

1.5

17. All of the questions had just two response options. Thus, if people answered every question and they were *guessing* on every question, we would expect them to get 50% of the questions correct. What percentage of people got exactly 5 of the 10 questions correct?

11

18. What percentage of people scored at or below chance (i.e., 5 correct responses out of 10) on this test? Use the cumulative percentage column to answer this question.

19. After taking standardized tests, you typically get a raw score as well as a **percentile rank**, *the percentage of scores a given score is higher than*. For example, if you scored at the 95th percentile, you would know that you scored as well as or better than 95% of the people who took the test. The same thing can be done for this data file by using the cumulative percent column of the frequency distribution table. What Science Knowledge test score is at the 95th (i.e., the 94.9th) percentile?

20. What Science Knowledge test score is at the 9th percentile?

21. What is the percentile rank for a Science Knowledge test score of 7?

22. What is the percentile rank for a Science Knowledge test score of 3?

23. On the GSS, people were asked how many years of school they completed. Before you graph the data in SPSS, what two responses to this question do you think will be the most common in the United States? Be sure that you can explain why you think these answers would be given frequently.
    a. 8
    b. 10
    c. 12
    d. 16

24. Use SPSS to create a bar graph of the YearsofEducation variable. What was the most frequently occurring response?

25. What percentage of the respondents completed exactly 12 years of school?

26. What percentage of the respondents completed 16 or more years of school?

27. What percentage of the respondents completed fewer than 12 years of school?

28. As you work with data throughout this course, you will find that frequency distribution graphs (i.e., bar graphs and histograms) can look quite different for different variables. Look at the graph for the Science Knowledge scores that you created earlier. This graph is very close to a bell-shaped curve, but it is a bit skewed. Is it positively skewed or negatively skewed?

neg

29. Use SPSS to generate a histogram for the number of hours people report watching TV each day. When you create a histogram in SPSS, you can click on a box that says "show normal curve on histogram." Select this option so that SPSS will also show a normal curve along with the data. Is the graph positively or negatively skewed?

pos

30. On the GSS, respondents were asked how old they were when their first child was born (variable is named AgeFirstChildBorn). Use SPSS to create a histogram for this variable. Do the data look to be positively skewed, negatively skewed, or bell shaped?

pos

31. As part of the GSS, respondents were given a vocabulary test that consisted of 10 words. Create a bar chart of the number correct on the vocabulary test (VocabTest). Do the data look to be positively skewed, negatively skewed, or bell shaped?

neg

32. Scores on the vocabulary test are not perfectly normally distributed. Is this distribution platykurtic or leptokurtic?

lep

33. If a test is relatively easy and most people get between 90% and 100%, but a few people get low scores (10%–20%), would that data be positively skewed or negatively skewed?

neg

MEASUREMENT

A researcher wonders if the age at which people have their first child is related to their level of education. To investigate this possibility, he divides people who have had children into four education brackets: high school or less, some college, college degree, and advanced degree. He then records the age at which each person had his or her first child. Circle the scale of measurement used for each of the variables listed below:

34. Education bracket: Nominal (Ordinal) Interval/ratio
35. Age at birth of first child: Nominal Ordinal (Interval/ratio)

A researcher designs a study in which participants with low levels of HDL (*high-density lipoprotein*) cholesterol are randomly assigned to take either a cholesterol-lowering drug or a placebo each day. After 3 months, HDL levels are measured. Women and men often react differently to treatments, and so the researcher also records the gender of each participant. Circle the scale of measurement used for each of the following variables:

36. Treatment (drug vs. placebo): (Nominal) Ordinal Interval/ratio
37. HDL cholesterol levels: Nominal Ordinal (Interval/ratio)
38. Gender of the participants: (Nominal) Ordinal Interval/ratio

A researcher used government records to classify each family into one of seven different income categories ranging from "below the poverty line" to "more than $1 million a year." The researcher used police records to determine the number of times each family was burglarized. Circle the scale of measurement used for each of the following variables:

39. Income category: Nominal (Ordinal) Interval/ratio
40. Number of times burglarized: Nominal Ordinal (Interval/ratio)

Determine which of the variables are discrete and which are continuous.

41. Number of times burglarized (Discrete) Continuous
42. Gender of the participants (Discrete) Continuous
43. Age at birth of first child Discrete (Continuous)

Chapter 1 Practice Test

1. A marriage counselor measures anger toward one's spouse by measuring the number of pins participants stick into a doll representing the participant's spouse. What is the scale of measurement for anger?
   a. Nominal
   b. Ordinal
   c. Interval/ratio

2. In the previous question, is this measure of anger continuous or discrete?
   a. Continuous
   b. Discrete

3. The marriage counselor wonders if men or women will display more anger and so he records the gender of each participant. What is the scale of measurement for gender?
   a. Nominal
   b. Ordinal
   c. Interval/ratio

4. A sports psychologist ranks swimmers according to their finishing place in a 100-m race (i.e., first, second, third, etc.). What scale of measurement was used for finishing place?
   a. Nominal
   b. Ordinal
   c. Interval/ratio

5. A sports psychologist asks swimmers to record the number of minutes they spend swimming each day. Is this measure of swimming continuous or discrete?
   a. Continuous
   b. Discrete

6. To measure pain tolerance, a researcher asks participants to submerge their arm in a bucket of water that is 34 °F and to keep it there as long as possible. The total time the participant kept his or her arm in the bucket of water was used as the measure of pain tolerance. What is the scale of measurement for this variable?
   a. Nominal
   b. Ordinal
   c. Interval/ratio

7. Is the total time the participant kept his or her arm in the bucket of water a continuous or a discrete variable?
   a. Continuous
   b. Discrete

8. When do researchers typically use bar graphs?
   a. When the data are continuous
   b. When the data are discrete
   c. When the data are interval/ratio
   d. When the data are ordinal

9. When do researchers typically use line graphs or histograms?
   a. When the data are continuous
   b. When the data are discrete
   c. When the data are interval/ratio
   d. When the data are ordinal
   e. When the data are nominal

10. Bridget received a score of 90 on an exam that put her at the 45th percentile. How did she do on this test?
    a. Her score of 90 is excellent; she scored better than 55% of the people in the class.
    b. She scored better than 45% of the people in the class.

11. On a chemistry exam, a few people received scores of 90% to 100%, but most received scores below 50%. How is this distribution skewed?
    a. Positively skewed
    b. Negatively skewed

12. A marriage counselor measures anger toward one's spouse by the number of pins stuck into a doll. Most people only stabbed the doll with 2 to 5 pins, but two people stabbed the doll with 52 pins. How is this distribution skewed?
    a. Positively skewed
    b. Negatively skewed

13. After completing a documentary film about the history of different font styles, the producer wonders if people will find the topic a bit boring. To measure boredom, she asks people to watch the film and then videotapes them while they are watching the film. Later, a research assistant watches the videos and records the number of times each person yawned. The data are at right.

X	*F*
0	1
1	2
2	1
3	1
4	2
5	6
6	7
7	6
8	5

How many people were videotaped while they watched the film?

a. 8
b. 31
c. 36

14. For the data described in Question 13, how many people yawned 7 times?

a. 6
b. 7
c. 8

15. For the data described in Question 13, what percentage of people did not yawn at all?

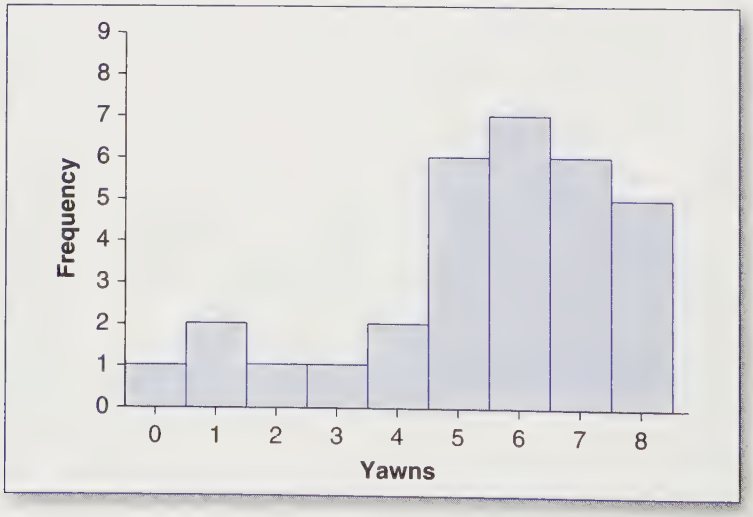
a. 1%
b. 3.2%
c. 0%
d. 1.2%

16. For the data described in Question 13, what percentage of the people yawned 4 or fewer times?

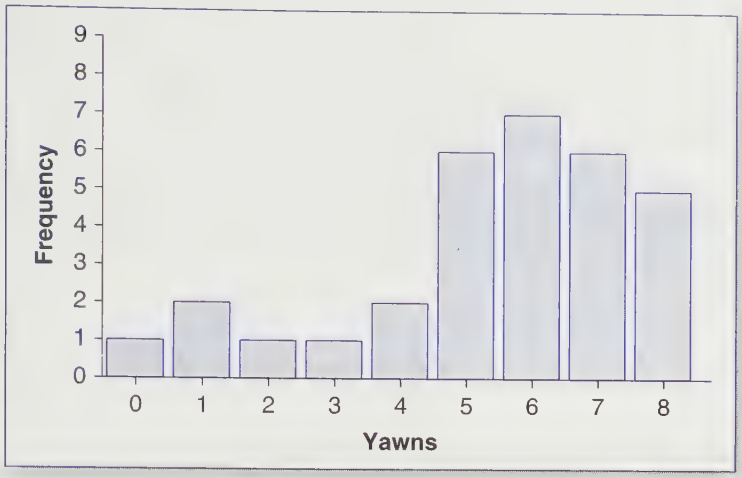
a. 6.45%
b. 16.13%
c. 7.52%
d. 22.58%

17. Which of the following is the best graph of the data described in Question 13?

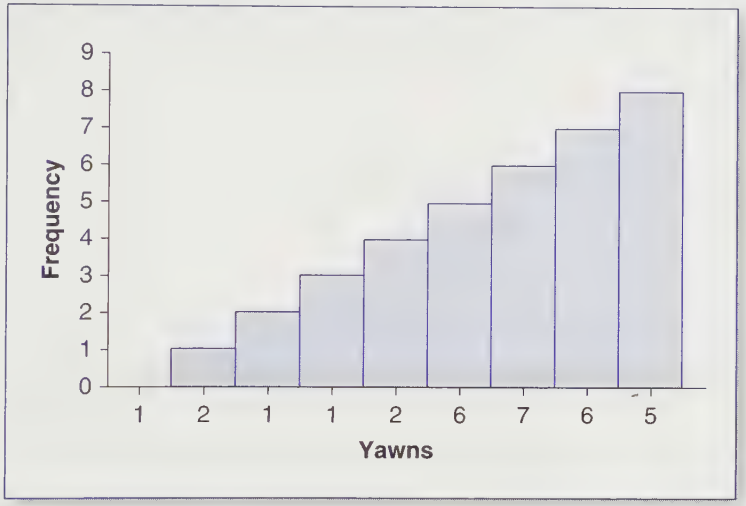
a.



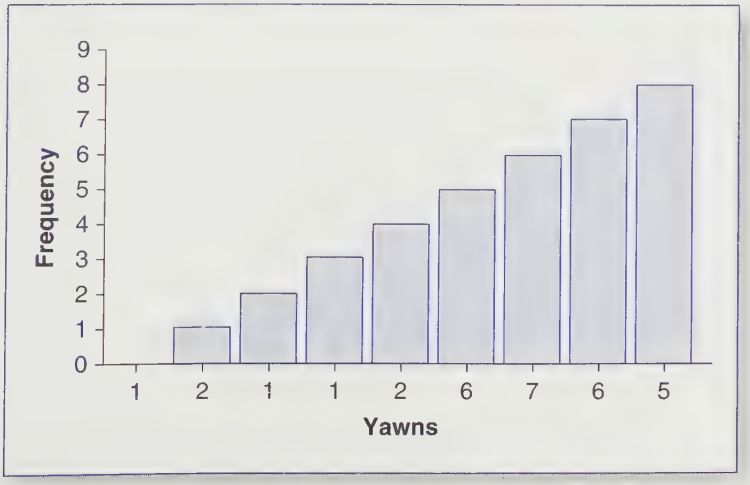
b.


Frequency
0
1
2
3
4
5
6
7
8
9
0 1 2 3 4 5 6 7 8
Yawns

c.


Frequency
0
1
2
3
4
5
6
7
8
9
1 2 1 1 2 6 7 6 5
Yawns

d.


Frequency
0
1
2
3
4
5
6
7
8
9
1 2 1 1 2 6 7 6 5
Yawns

18. What is sampling error?
    a. The difference between qualitative and quantitative data
    b. The difference between a sample statistic and a population parameter
    c. The difference between an inferential statistic and a descriptive statistic

19. A polling organization asked a representative sample of 50- to 55-year-olds living in the United States to determine how much they had saved for retirement. The average amount saved was $125,000. Which of the following best describes this number?
    a. Sample parameter
    b. Sample statistic
    c. Population statistic
    d. Population parameter

20. A polling organization asked a representative sample of 50- to 55-year-olds living in the United States to determine how much they had saved for retirement. The average amount saved was $125,000. The pollsters use this information to make the argument that people in the United States are not saving enough for retirement. This is an example of a
    a. descriptive statistic.
    b. inferential statistic.

21. Which of the following data files shows the correct way to enter the heights of six people into SPSS?

a.

VAR00001	VAR00002	VAR00003	VAR00004	VAR00005	VAR00006
68.00	55.00	69.00	71.00	54.00	63.00

b.

VAR00001
68.00
55.00
69.00
71.00
54.00
63.00

REFERENCES

American Psychological Association. (2010). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Carifio, J., & Perla, R. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of the Social Sciences, 3,* 106–116.

Field, T. (2010). Touch for socioemotional and physical well-being: A review. *Developmental Review, 30*(4), 367–383.

Field, A. P. (2013). *Discovering statistics using IBM SPSS Statistics: And sex and drugs and rock 'n' roll* (4th ed.). London: Sage.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law, 2*(2), 293–323.

Harris, G. T., Rice, M. E., & Quinsey, V. L. (1993). Violent recidivism of mentally disordered offenders: The development of a statistical prediction instrument. *Criminal Justice and Behavior, 20,* 315–335.

Rosa, L., Rosa, E., Sarner, L., & Barrett, S. (1998). A close look at therapeutic touch. *Journal of the American Medical Association, 279*(13), 1005–1010.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677–680.

Yang, M., Wong, S. P., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin, 136*(5), 740–767.

CHAPTER 2

Central Tendency

LEARNING OBJECTIVES

After reading this chapter, you should be able to do the following:

- Compute and interpret the mean, the median, and the mode
- Identify when to use the mean, the median, or the mode when describing a distribution's central tendency

CENTRAL TENDENCY

You are probably already familiar with the notion of **central tendency**. For example, if your five history exam scores for a semester were 33%, 81%, 86%, 96%, and 96%, the "center" of these scores summarizes your academic performance in history. However, did you know that your history instructor could use the mean, the median, or the mode to find the center of your scores? If your instructor used the *arithmetic average* (i.e., the **mean**, 78.4%), you would get a C. Although the mean is the most common measure of central tendency, there are other options. She could use the *middle test score* (the **median**, 86%), and you would get a B. She could also use the *most frequently occurring test score* (the **mode**, 96%), and you would get an A. Clearly, the mode gives the most favorable picture of your performance, but which measure of center gives the most accurate picture? Researchers frequently must summarize large data sets by presenting a single measure of their center, and every time they do so, they must determine whether the mean, the median, or the mode offers the most accurate summary of the data. It should be clear from the history exam example that one's choice of mean, median, or mode can dramatically change the interpretation of the data. Luckily, there are rules of thumb that help you decide when to use each of these measures of center.

The mean is the arithmetic average (the sum of the scores divided by the number of scores). While the mean is the most common measure of a data set's center, there are situations when it cannot be used. For example, *when the data being summarized are nominal, the mean cannot be used.* Suppose we recorded the academic majors of students in a class and 17 responded psychology, 9 nursing, 7 sociology, and 8 social work. These academic majors are categories, not values; one cannot compute the average academic major. In other words, academic major is a nominal variable. In this case, psychology is the most frequently reported academic major, and therefore, it is the center of the data. *When the data are nominal (i.e., when the data are categories rather than values), you must use the mode to summarize the center.*

Reading Question

1. What measure of central tendency is used when data are measured on a nominal scale?
   a. Mean
   b. Median
   c. Mode

While the mode is the only option when data are nominal, *the median is the best option when data are ordinal.* The median of a distribution of scores is the value at the 50th percentile, which means that half of the scores are below this value and half are above it. For example, suppose we wanted to know the center class rank of students in a statistics course with five freshmen, seven sophomores, four juniors, and three seniors. We could assign ordinal positions to each class rank so that freshman = 1, sophomore = 2, junior = 3, and senior = 4. This would result in the following distribution of scores: 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4. Although we used numbers to represent the ordinal positions, it would not make mathematical sense to compute the mean of class rank (i.e., freshmen, sophomores, juniors, seniors). When a variable is ordinal, the best measure of center is the median. *After the scores are arranged from lowest to highest,* the median is the score with exactly the same number of scores above it and below it. For example, after the 19 class ranks are arranged from lowest to highest, the 10th highest score would have nine scores above it and nine scores below it. The 10th highest score is a 2: 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4. Thus, the median value, or center value, is sophomore. If a data set has an even number of scores, you find the median by computing the average of the middle two scores.

Reading Question

2. What is the median for this set of scores? 3, 3, 3, 4, 4, 5, 5, 5
   a. 3
   b. 4
   c. 5

Reading Question

3. What measure of central tendency should be used with ordinal data?
   a. Mean
   b. Median
   c. Mode

When working with interval or ratio data, you need to choose between the mean and the median to represent the center. In general, you should use the mean to summarize interval or ratio data; however, if the data set contains one or more "extreme" scores that are very different from the other scores, you should use the median. For example, if you wanted to summarize how many text messages you typically send/receive during a typical hour-long class, you could use the mean or the median because the number of texts sent/received in a class is ratio data. Suppose that during your last seven classes, you sent/received the following numbers of messages: 7, 9, 38, 6, 7, 8, and 7. The mean number of text messages would be 11.7, but this is not really a typical value for you. During the vast majority of classes (i.e., six of your last seven classes), the number of messages you sent was less than 11.7. The mean does not represent the center of these data very well because a single extreme score is inflating the mean, "pulling" it away from the center. Statisticians would consider the 38 value an **outlier** because it is *a very extreme score compared with the rest of the scores in the distribution.* The median is far less affected by extreme scores. The median would be 7 (6, 7, 7, 7, 8, 9, 38), and this would be a much better summary of your in-class texting habits. When to use each measure of central tendency is summarized in Figure 2.1.

Figure 2.1 When to Use Measures of Central Tendency

Measure of central tendency	*When to use the measure*
Mean	With interval/ratio data that are normally distributed; no outliers
Median	With ordinal data With interval/ratio data that are skewed or have outliers
Mode	With nominal data

Reading Question

4. What measure of central tendency is obtained by adding all the scores and then dividing by the number of scores?
   a. Mean
   b. Median
   c. Mode

Reading Question

5. What measure of central tendency is the value that has half of the scores above it and half of the scores below it?
   a. Mean
   b. Median
   c. Mode

Reading Question

6. What measure of central tendency should be used when the data are interval and there are extreme scores in the distribution?
   a. Mean
   b. Median
   c. Mode

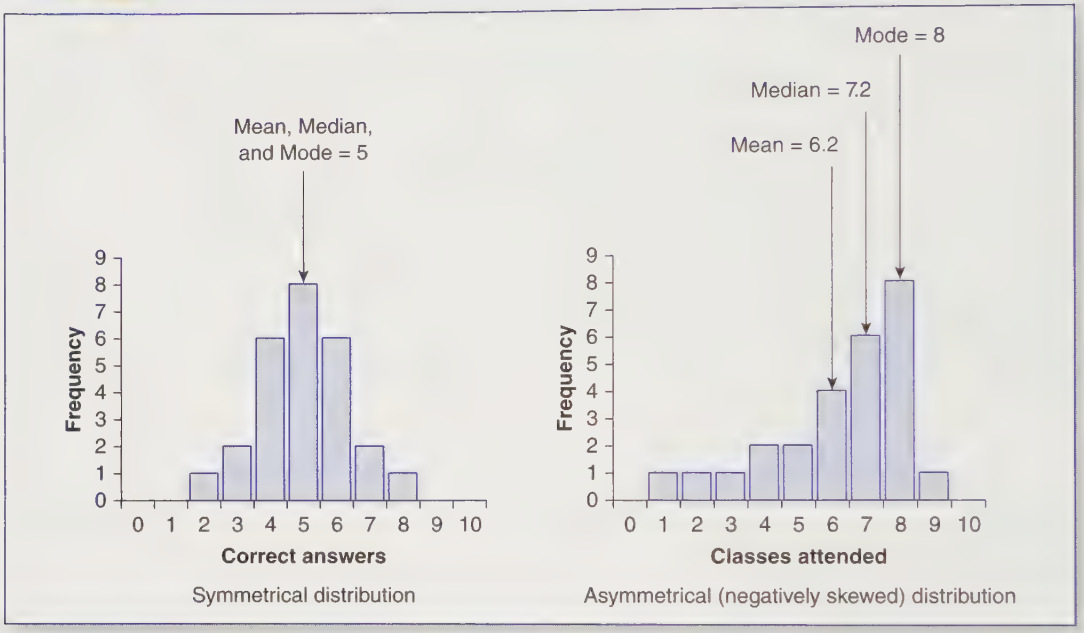
Reading Question

7. Extreme scores are also called
   a. outliers.
   b. modes scores.

Another situation in which you should use the median even though you have interval or ratio data is when the distribution is asymmetrical, or skewed. A distribution of scores is skewed if the "tail" on one side of the distribution is substantially longer than the tail on the other side. The bar graphs in Figure 2.2 illustrate a symmetrical distribution and a negatively skewed distribution, respectively. The mean should be used for symmetrical distributions, and the median should be used for highly skewed distributions. It is worth noting that when a distribution is symmetrical (or close to being symmetrical), the mean, the median, and the mode

are all very similar in value. However, when a distribution is very asymmetrical, the mean, the median, and the mode are different. In asymmetrical distributions, the mean is "pulled" toward the distribution's longer tail. This fact is illustrated by the asymmetrical graph in Figure 2.2.

Figure 2.2 A Bar Graph of Symmetrical Data (Correct Answers) and a Bar Graph of Negatively Skewed Data (Classes Attended)



Reading Question

8. What measure of central tendency should be used when a distribution of scores measured on the interval or ratio scale of measurement is skewed?
   a. Mean
   b. Median
   c. Mode

Reading Question

9. When a distribution of scores is skewed, the median and the mean will be similar.
   a. True
   b. False

COMPUTING THE MEAN

To compute the mean of the following sample of scores, you would add up the scores and then divide by the number of scores. If you do these calculations, you will find that the mean is 73:

100, 70, 80, 90, 50, 60, 70, 80, 90, 40

Although this is probably something that you already know how to do without referring to a formula, you will need to be familiar with the following notation to understand the more complicated statistical formulas covered in later chapters. The formula for computing the mean (M) of sample data is

$$M = \frac{\sum X}{N}.$$

The numerator of the formula ($\sum X$), read as "sum of X," is using statistical notation to indicate that the scores (i.e., Xs) should be added. The sigma ($\sum$) tells you to sum, or add, whatever comes after it, so $\sum X$ means sum all the Xs. The N in the denominator of the formula refers to the number of scores. So, the **statistical formula** for the mean is literally a set of instructions telling you to add the scores and then divide them by the number of scores. In this case, the sum of the Xs ($\sum X$) is 730, the N is 10, and therefore, the mean is 73:

$$M = \frac{\sum X}{N} = \frac{100 + 70 + 80 + 90 + 50 + 60 + 70 + 80 + 90 + 40}{10} = \frac{730}{10} = 73.$$

In subsequent chapters, you will be learning statistical procedures that are more complicated than the mean. Understanding and computing these more advanced statistics will be much easier if you learn to read statistical notations (e.g., $\sum X$). In the long run, learning to read statistical formulas will be much easier than trying to memorize the exact order of multiple mathematical operations.

Reading Question

10. What does $\sum X$ tell you to do?
    a. Sum the scores (Xs)
    b. Compute the mean

Reading Question

11. What does N represent?
    a. The number of scores
    b. Measurement categories

Reading Question

12. A statistical formula is
    a. a helpful set of instructions indicating how to compute something.
    b. a bunch of meaningless symbols I should skip when I'm reading.

You will also need to be able to compute the mean from data presented in a frequency distribution table. For example, the test scores listed earlier could be presented in a frequency distribution table, as in Table 2.1.

The frequency distribution indicates that there is one person with a score of 100, two people with scores of 90, two people with scores of 80, and so on. To obtain the sum of the scores, you could add each individual number, like you did previously, but with larger data sets, it is more

Table 2.1 Frequency Distribution Table of Variable Scores

X	f
100	1
90	2
80	2
70	2
60	1
50	1
40	1

efficient to work with the data as they are presented in the table. There is just one person with a score of 100, and so we will include just one score of 100. However, there are two people with a score of 90. Rather than put two 90s into the equation, you can multiply 90 by 2. More generally, you will need to multiply each score by the number of people who had that score ($\sum(Xf)$). This is illustrated below:

$$\sum X = 100(1) + 90(2) + 80(2) + 70(2) + 60(1) + 50(1) + 40(1) = 730.$$

Table 2.2 Frequency Distribution Table for a Larger Set of Scores

X	f
100	5
90	7
80	8
70	10
60	6
50	9
40	3

Of course, the sum of scores is identical to what you computed earlier. As before, you must divide the sum of scores by the number of scores (N) to find the mean. The N, when data are presented in a frequency table, is the sum of the frequencies ($\sum f$). In this case, $N = \sum f = 10$, and so the mean is computed as

$$M = \frac{\sum X}{N} = \frac{730}{10} = 73.$$

Knowing how to incorporate the frequencies, f, when computing the mean for data in a frequency table can save you time when working with larger data sets. For example, suppose that you had test scores from a larger sample of people (Table 2.2).

You could find the mean of the scores in this frequency table in the following way:

$$M = \frac{\sum X}{N} = \frac{100(5) + 90(7) + 80(8) + 70(10) + 60(6) + 50(9) + 40(3)}{48}$$
$$= \frac{3{,}400}{48} = 70.83$$

First you multiply each score by its frequency, then you add all those values together, and finally you divide by N. It is important to remember that N when working with a data set from a frequency table is always the sum of the frequencies ($\sum f$), *not* the number of scores listed in the table. So, in this example, N is 48, not 7.

Reading Question

13. Which of the following is the best way to compute the mean for the following data?

X	f
3	4
2	7
1	5

a. $M = (3(4) + 2(7) + 1(5))/16$
b. $M = (3 + 2 + 1)/3$
c. $M = ((3 + 2 + 1) + (4 + 7 + 5))/6$

The data in the previous problems came from samples. The computations are identical when you are working with a population. As with the sample, you will sum the scores and divide by N. However, the notation is a bit different. Greek letters are used to represent population parameters (i.e., μ, pronounced "myoo," represents a population mean), while Arabic letters are used to represent sample statistics (i.e., M represents a sample mean). The formula for the population mean (μ) is provided below, but you should note that it is computed in exactly the same way as the sample mean (M):

$$\mu = \frac{\sum X}{N}.$$

Reading Question

14. The sample mean is represented by
a. M.
b. μ.
c. both M and μ.

Reading Question

15. The population mean is represented by
a. M.
b. μ.
c. both M and μ.

FIND THE MEDIAN

The median is the midpoint of a distribution of scores. When working with a list of scores, you begin by putting the scores in order from lowest to highest (or highest to lowest). For example, the test scores 100, 70, 80, 90, 50, 60, 70, 80, 90, and 40 must be reordered as follows:

40, 50, 60, 70, 70, 80, 80, 90, 90, 100

The median is the middle score in the list. In this case, there are an even number of values, and so there is not one middle score but two. The two middle scores are 70 and 80. To find the median, compute the average of the two middle scores. Thus, the median would be 75. When the number of scores is odd, the median is the one value that is the exact center of the *ordered* list. For example, for the following list of scores, the median would be 6:

7, 7, 6, 6, **6**, 5, 4, 3, 1

When working with data from a frequency table, the scores will already be in order. But you will need to determine how many scores from either end of the list you need to "move in" to find the middle score. In Table 2.3, there are 48 scores ($N = \sum f = 48$), and so the median is between the 24th and 25th scores from either end of the list. If you start from the bottom, you will see that there are

three people with scores of 40, nine with 50s, and six with 60s. So far, we have counted up 18 scores and we need to count up to the 24th and the 25th scores, which are 6 and 7 more scores up the order, respectively. The next category of scores, 70, has 10 scores in it. Therefore, the 24th and the 25th scores must both be 70, and the median is the average of these two 70s, which is 70.

Table 2.3 Frequency Distribution Table for a Larger Set of Scores With Median Identified

X	*f*
100	5
90	7
80	8
70	10
60	6
50	9
40	3

The 24th and 25th scores are in this group of 10 scores.

Reading Question

16. Find the median of this list of scores: 5, 6, 4, 7, 8. [Hint: What is the first step to finding the median?]

a. 4

b. 5

c. 6

Reading Question

17. Find the median of this frequency table of scores:

X	*f*
3	4
2	7
1	5
0	1

a. 1

b. 2

c. 3

FIND THE MODE

The mode is the most frequently occurring score in a distribution. To locate the mode in the frequency distribution table, you look for the measurement category (X value) with the highest frequency. In Table 2.4, the mode would be 70 because 10 people had scores of 70.

Table 2.4 Frequency Distribution Table for a Larger Set of Scores, With the Mode Bolded

X	f
100	5
90	7
80	8
70	**10**
60	6
50	9
40	3

Reading Question

18. Find the mode of this set of scores:

X	f
3	4
2	7
1	5
0	1

a. 1
b. 2
c. 3

SPSS

Data File

To compute measures of central tendency using SPSS, you will need to begin by entering the data. As described in Chapter 1, all the scores for a particular variable should be entered in one column. *You cannot*

enter a frequency distribution table into SPSS; instead you must enter individual scores. For example, to enter the scores from Table 2.5, you would need to enter 100 once, 90 twice, and so forth.

When you are done, your data file should look like the one in Figure 2.3. The exact order of the values is not important, but you should be sure that all 10 scores are entered in a single column.

Reading Question

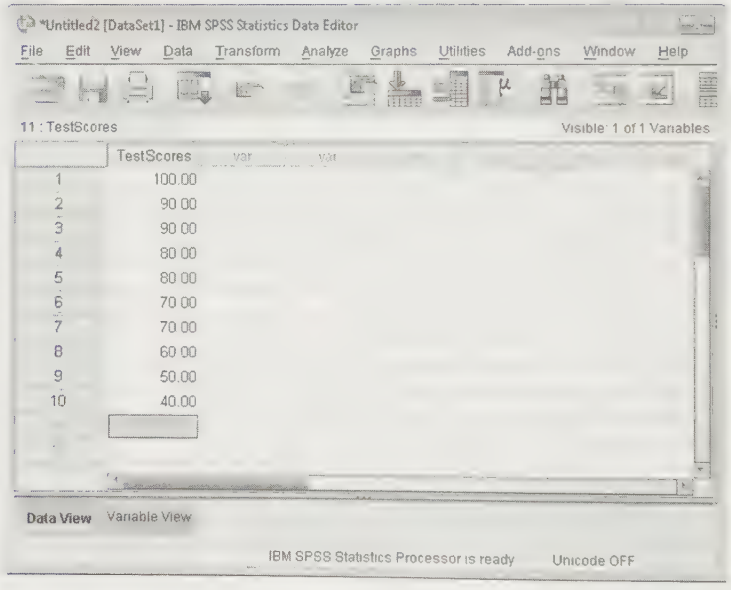
19. When entering data into SPSS, you can enter a frequency table of the data; you do not have to enter each score individually.

a. True
b. False

Table 2.5 Frequency Distribution Table of the Variable Called Scores

X	f
100	1
90	2
80	2
70	2
60	1
50	1
40	1

Figure 2.3 SPSS Screenshot of the Data Entry Screen of the Variable Labeled Test Scores



Obtaining Measures of Central Tendency Using SPSS

Do the following to generate measures of central tendency using SPSS:

- Click on the Analyze menu. Choose Descriptive Statistics and then Frequencies (see Figure 2.4).
- Move the variable(s) of interest into the Variable(s) box (see Figure 2.5).

Figure 2.4 SPSS Screenshot of the Analyze Menu for Descriptive Statistics

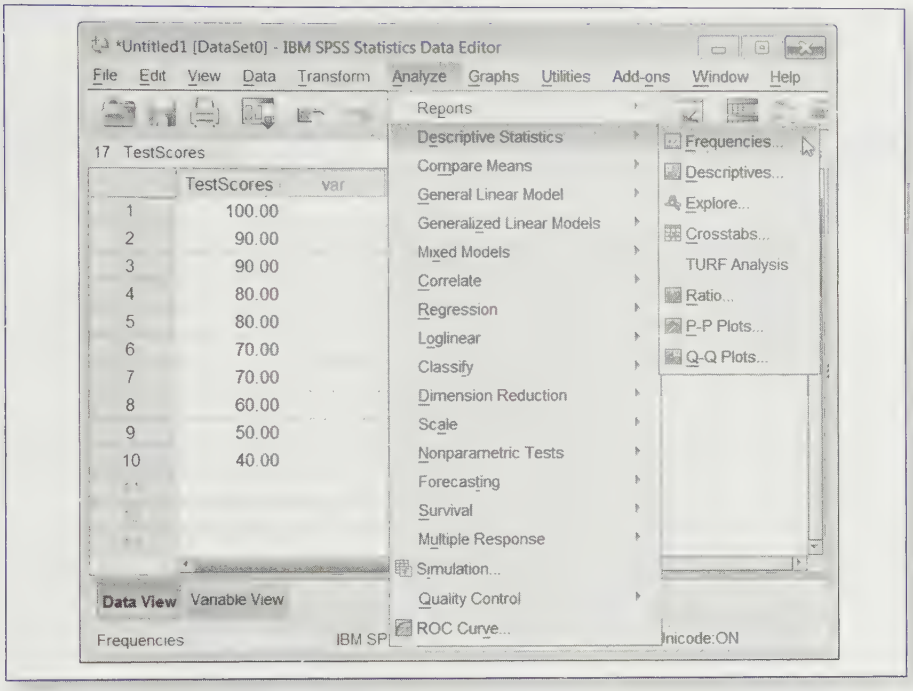
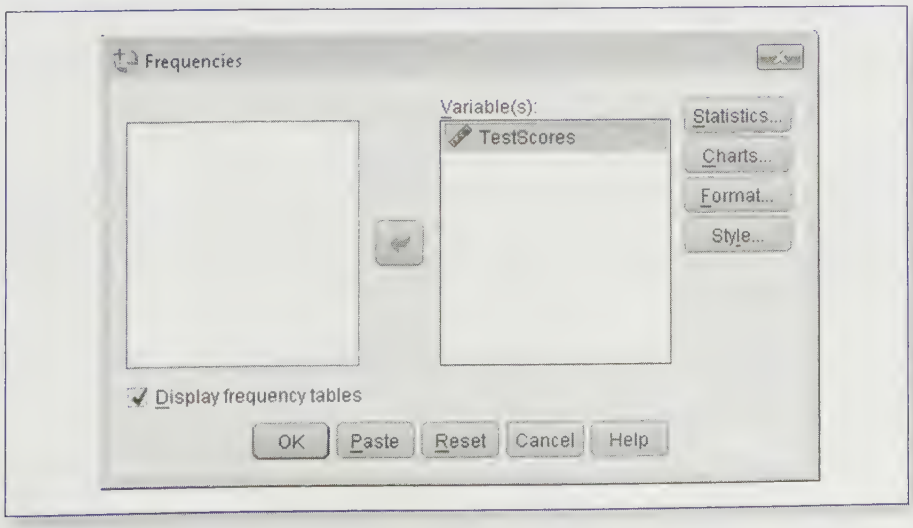
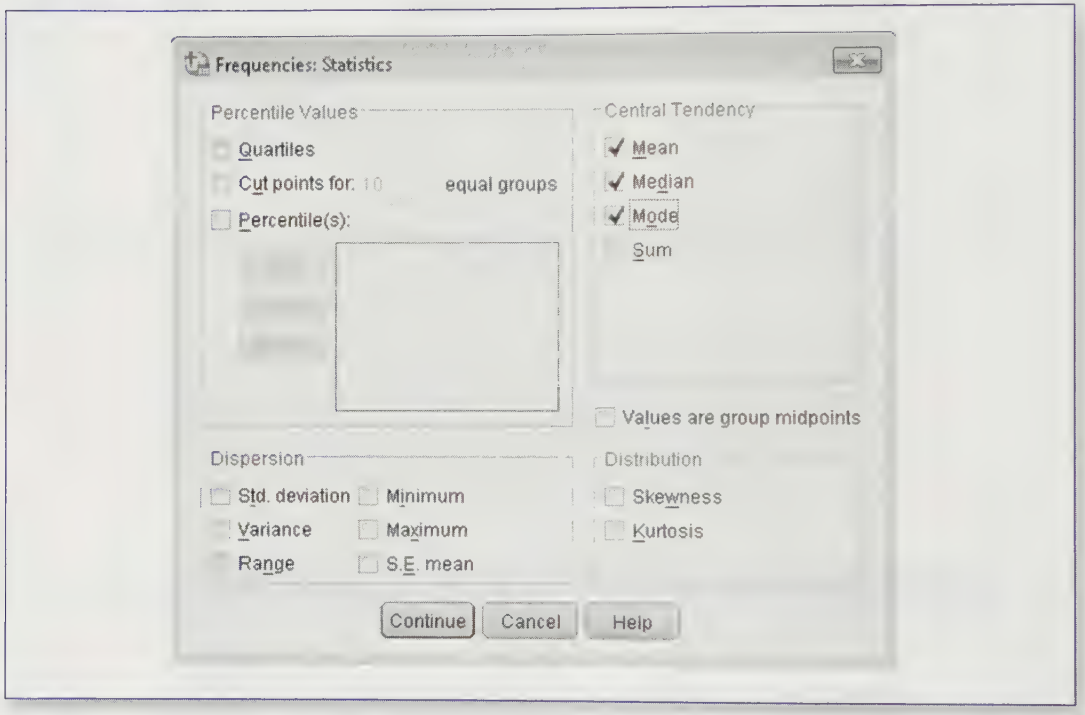


Figure 2.5 SPSS Screenshot of Choosing the Variables for Descriptive Statistics



- Make sure the Display Frequency Tables box is checked if you want a frequency distribution table. Uncheck the box if you do not want a frequency table.
- Click on the Statistics button.
- Click on the boxes for mean, median, and mode, and then click on the Continue button (see Figure 2.6).
- Click OK.

Figure 2.6 SPSS Screenshot of Choosing Measures of Central Tendency



Output

Your output file should look similar to that in Figure 2.7. Note that this data file had multiple modes because three different test scores each had a frequency of 2. SPSS only tells you one mode, but it does make a note that there are additional modes. To find the other modes, you need to look at the frequency distribution table.

Reading Question

20. What was the mean of these data?
    a. 73.0
    b. 75
    c. 70

Figure 2.7 SPSS Output for the Central Tendency of the Variable Score

Statistics

testscores

N	Valid	10
	Missing	0
Mean		73.0000
Median		75.0000
Mode		70.00[a]

a. Multiple modes exist. The smallest value is shown

testscores

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	40.00	1	10.0	10.0	10.0
	50.00	1	10.0	10.0	20.0
	60.00	1	10.0	10.0	30.0
	70.00	2	20.0	20.0	50.0
	80.00	2	20.0	20.0	70.0
	90.00	2	20.0	20.0	90.0
	100.00	1	10.0	10.0	100.0
	Total	10	100.0	100.0	

OVERVIEW OF THE ACTIVITY

In Activity 2.1, you will compute measures of central tendency (mean, median, and mode) by hand and using SPSS. You will practice determining which measure of central tendency is appropriate for different types of data. Finally, deviation scores will be introduced because (1) they will help you understand the mean and (2) the mean and deviation scores will help you understand an important concept in the next chapter, the standard deviation.

Activity 2.1: Central Tendency

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Compute the mean, the median, and the mode for data presented in a frequency table
- Identify whether the mean, the median, or the mode best represents the center of a given distribution of data
- Recognize from a graph when data will produce a misleading mean, median, or mode
- Explain how the mean perfectly balances positive and negative deviations scores

PART I: COMPUTING THE MEAN, THE MEDIAN, AND THE MODE

You asked 10 fellow statistics students how many courses they were taking this semester. The 10 people gave you the following answers: 5, 4, 4, 4, 3, 3, 3, 2, 2, and 1.

1. Create a frequency table of the "Number of Courses" data.

5 3
4 3
3 2
2 1
1

2. Create a frequency bar graph of the "Number of Courses" data.

3. Compute the mean for these data.

3.1

4. Find the median for these data.

3

5. Find the mode for these data.

both 3 & 4

6. Which measure of central tendency would be the best in this situation?

   a. Mean
   b. Median
   c. Mode

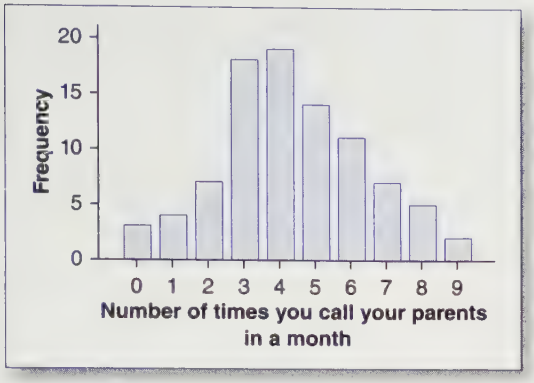
You should have found that the mean and the median were 3.1 and 3, respectively. There were two modes, 3 and 4. When it comes to choosing the best measure of central tendency, some statisticians would say that the number of courses is discrete and therefore the mean should not be used. They would argue that the value of 3.1 is not a good measure of central tendency because no one can ever takes 3.1 courses. However, other statisticians would argue that as long as you remember that no one actually takes 3.1 courses, the mean of 3.1 is a more accurate measure of the center than 3. In this book, we will use the mean as a measure of central tendency for both continuous and discrete variables as long as (1) the variable is measured on an interval/ratio scale, (2) there are no outliers, and (3) the distribution is "basically" symmetrical (i.e., not "very" skewed).

PART II: CENTRAL TENDENCY GRAPH EXERCISE

For each of the following situations, determine if you should use the mean, the median, or the mode as the measure of central tendency.

7. The variable is nominal. Mean Median **Mode**
8. The variable is ordinal. Mean **Median** Mode
9. The variable is interval/ratio. **Mean** Median Mode
10. The variable is interval/ratio but is highly skewed. Mean **Median** Mode
11. The variable is interval/ratio, but there are outliers. Mean **Median** Mode

For the following problems, determine which measure of central tendency is the most appropriate. Be sure to consider the scale of measurement as well as the shape of the distribution.



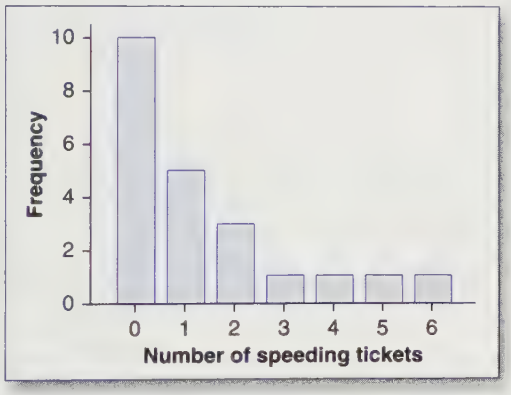
12. These data came from a sample of college students who attend college at least 100 miles away from their parents. They were asked how many times they called their parents in a typical month. The mean, the median, and the mode were 4.34, 4, and 4, respectively. Which measure of central tendency should be used, and why?

a. The mean because the data are interval/ratio and the distribution is not very skewed.

b. The mode because the data are discrete.

c. The median because the data are ordinal.

d. All measures of central tendency are equally appropriate for these data.

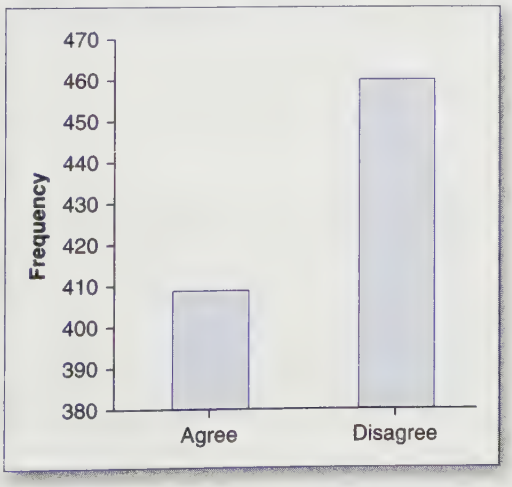


13. These data came from people attending a stock car race in Indiana. They were asked how many speeding tickets they had received in the previous 2 years. The mean, the median, and the mode were 1.45, 1, and 0, respectively. Which measure of central tendency should be used, and why?

a. The median because the data are ordinal.

b. The mean because the data are interval/ratio.

c. The median because the data are interval/ratio but skewed.

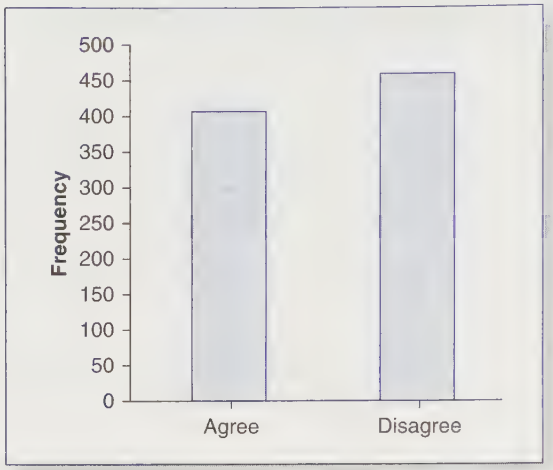


14. These data came from clinical psychologists who were attending a professional conference. They responded to the question, "Do you agree that patients' memories of past events are improved by hypnosis?" What is the best measure of central tendency for these data, and why?

a. The mode because the data are ordinal.

b. The median because the data are skewed.

c. The mode because the data are nominal.



15. This graph presents the same data that were presented in the preceding graph. Why do the graphs look so different?

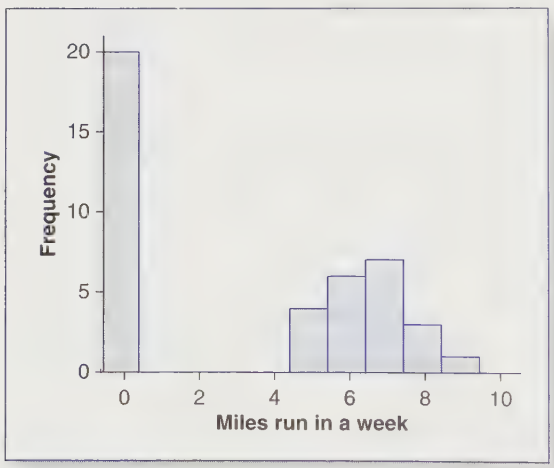
a. The range of values on the *y*-axis is different in the two graphs.

b. The second graph uses an interval/ratio scale while the first uses a nominal scale.

16. Which do you think is a more accurate representation of the data, and why? (It is important to note that memories are *not* improved by hypnosis.)

a. The first graph is better because it highlights the difference between the agree and disagree responses.

b. The second graph is better because it shows the true range of values.



17. These data were obtained from a sample of people in the supermarket. They were asked how many miles they ran in the previous week. The mean, the median, and the mode were 3.37, 5, and 0, respectively. In this case, no one number accurately describes the distribution of scores. In fact, one could argue that presenting only a measure of central tendency would be misleading. Which of the following is the best summary of these data?

a. The mean number of miles respondents reported running was 3.37 with a median of 5 and a mode of 0.

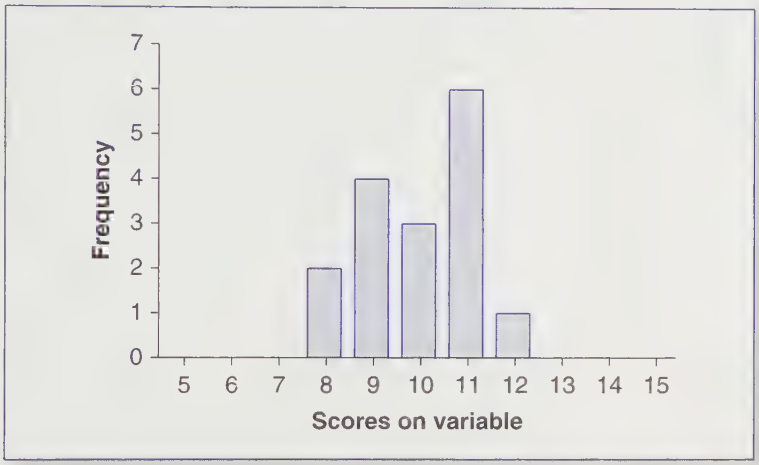
b. Most respondents reported that they did not run that week. Of those who did run, the reported miles ranged between 5 and 9, with a mean of 6.57.

c. The number of miles people reported running varied between 0 and 9, with a mean of 3.37 and a median of 5.

PART III: UNDERSTANDING THE MEAN

Part IIIA: Using Graphs to Understand the Mean

The mean, the median, and the mode are all used to represent the center of distributions of scores. However, the preceding questions made it clear that each of these statistics defines *center* differently. The mode defines the center as the most common score. The median defines the center as the middle score. Both of these definitions of *center* are easy to understand. The mean defines the center in a more sophisticated way. In this activity, we are going to use frequency bar graphs to explain how the mean defines the center of scores.



18. How many scores are in this distribution (i.e., What is *N*?)? 16

19. Compute the mean of this distribution by adding up all the scores (ΣX) and dividing by the number of scores (N). $M = \frac{\Sigma X}{N}$

10

You should have found $\Sigma X = 160$, $N = 16$, so $M = 10$. Make sure you understand how each of these values was computed.

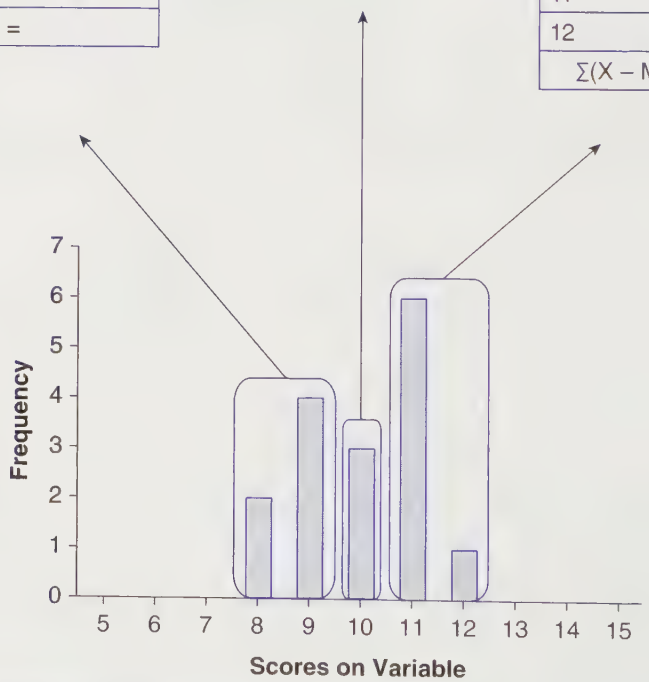
20. Although you have been able to compute the mean for a long time now, you probably don't know "what the mean does" or how it "defines" center. The mean defines center by balancing **deviation scores.** In other words, *the mean is the only value that perfectly balances all the positive and negative deviation scores in the distribution.* What is a deviation score? Every value in a distribution is a certain distance from the mean; this distance from the mean is the value's

deviation score. In this case, we are working with a sample, and so the deviation score is computed as $(X - M)$. For a population, the computations are the same, but the notation for the mean is μ rather than M; therefore, the deviation score notation when dealing with population is $(X - \mu)$. To understand how the mean balances deviation scores, compute each value's deviation score by using the formula $(X - M)$. For example, all scores of 8 have a deviation score of $(X - M) = (8 - 10) = -2$. The following graph has three boxes: one for the six scores less than the mean, a second for the three scores at the mean, and a third for the seven scores greater than the mean. Compute the deviation scores for each score and put them in the tables below. After you have computed the deviation scores, sum the deviations ($\sum(X - M)$) that are above, below, and at the mean.

Scores Below the Mean	
Score	Deviation $(X - M)$
8	
8	
9	
9	
9	
9	
$\sum(X - M) =$	

Scores At the Mean	
Score	Deviation $(X - M)$
10	
10	
10	
$\sum(X - M) =$	

Scores Above the Mean	
Score	Deviation $(X - M)$
11	
11	
11	
11	
11	
11	
12	
$\sum(X - M) =$	



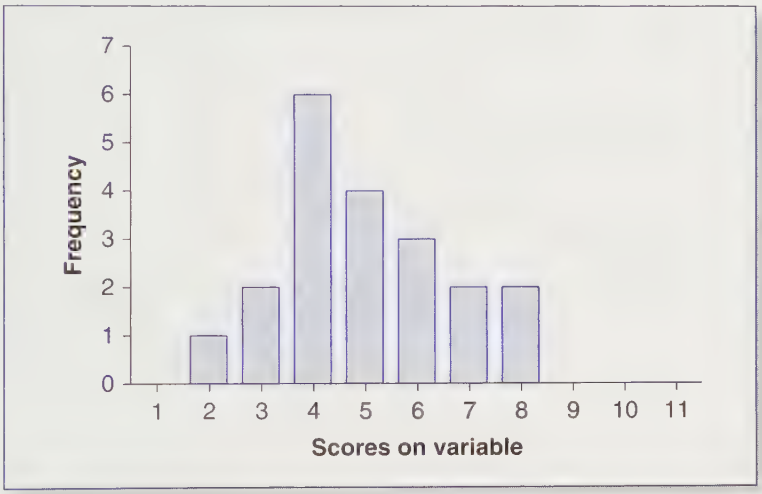
21. Record the sums of the positive and negative deviation scores below.

The sum for all negative deviation scores = −8

The sum for all positive deviation scores = +8

22. Now you should be able to see how the mean balances positive and negative deviation scores. The sum of all the positive deviations and the sum of all the negative deviations will always balance. The mean is the only value that perfectly balances all of the deviations scores. Therefore, if this always happens, the sum of all the deviations scores, $\Sigma(X - M)$, will always equal what value exactly? 0

23. Does this really always happen? Below is a completely different distribution of scores. Compute the mean of these scores. Remember, this is a frequency distribution so you have one score of 2, two scores of 3, and so on.



$$M = \frac{\Sigma X}{N}$$

You should have found $\Sigma X = 100$, $N = 20$, so $M = 5$. Make sure you understand how each of these values was computed.

5

24. Now that you have the mean of the above distribution of scores, compute the deviation scores for all of the scores less than the mean.

25. What is the sum of all the negative deviations (i.e., 2s, 3s, and 4s)? −13

26. Compute the deviation scores for all of the scores greater than the mean.

27. What is the sum of all the positive deviations (i.e., 6s, 7s, and 8s)? 13

28. Finally, even when we use a completely different set of numbers, the sum of all the deviation scores, $\sum(X - M)$, is equal to 0. This will always happen!

29. Which of the following statements about the mean is true? (Choose all that apply.)

    a. The mean is ALWAYS the exact point at which the sum of positive and negative deviation scores balances out.

    b. The mean is ALWAYS the score with the highest frequency.

    c. The mean is ALWAYS equal to the median.

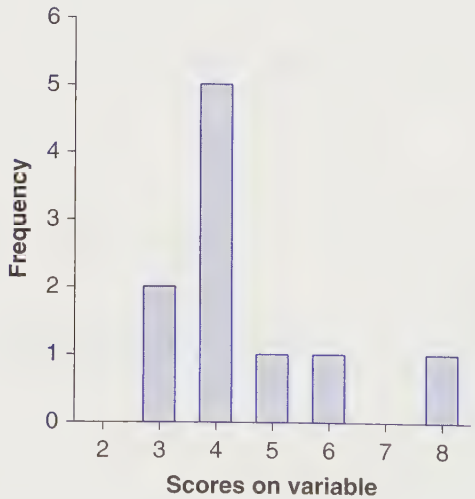
    d. The mean is ALWAYS equal to the mode.

    e. The mean is NEVER equal to the median.

    f. The mean is NEVER equal to the mode.

Part IIIB: Create Frequency Distributions

30. Below is a frequency graph with 10 scores on the left and an incomplete frequency table on the right. Fill in the frequency table based on the scores in the frequency distribution.



X (Score)	Frequency
2	
3	
4	
5	
6	
7	
8	

In the two displays above, the frequency graph and the frequency table convey similar information. You should become comfortable computing the mean, the median, or the mode from both of these displays. The mode is 4, the median is 4, and the mean is 4.5. Practice computing the mean using the STAT mode of your calculator. Don't move on to the next problem until you are sure you can do this. If you don't know how to use the STAT mode of your calculator, you can find instructions online by searching for your calculator.

Part IIIC: Computing Central Tendency From Frequency Distributions

31. Find the mode(s) for the data in the frequency distribution table below.
32. Find the median for the data in the frequency distribution table below.
33. Find the mean for the data in the frequency distribution table below.

X *(Score)*	f *(Frequency)*
2	1
3	1
4	3
5	2
6	3
7	0
8	0

$$M = \frac{\sum X}{N} =$$

Part IIID: Compute Deviation Scores

A deviation score represents the distance and direction by which an individual score varies from the mean. It is the number of units to the left or right of the mean that each score is located. Deviations to the left of the mean are negative, and deviations to the right are positive. [Note: Deviation score = $(X - M)$ or $(X - \mu)$.]

34. List all of the scores from the frequency table in the chart below. For example, there are three 6s in the table, so you need to list 6 three times, each in its own row, in the chart. Then, compute the deviation score for each of these values. Remember that you already computed the mean for all 10 of these scores in Question 33.

Score	Deviation Scores (X – M)

35. Sum all of the deviation scores. What value do you get? _____

PART IV: SUMMARY POINTS

36. One of the key points of this activity is that the sum of the distances from the mean (i.e., the deviation scores) ALWAYS equals 0. The positive and negative deviations will ALWAYS balance out. This balancing of *distances* will ALWAYS occur even when the *number* of positive and negative deviations is unequal. Which of the following statements is always true about the mean? (Choose all that apply.)
    a. The mean will always be the middle score; it will have the same number of scores above it and below it.
    b. The mean will always be the most common score in a distribution.
    c. The mean is the only value that exactly balances the distances of the deviations to the left and right of the mean.

37. The deviation scores for all distributions of scores always sum to zero. However, distributions of scores can vary quite a bit with regards to the typical distance scores are from the mean (deviation score). Which of the following distributions would have a greater typical deviation score?
    a. A class where everyone is between the ages of 20 and 21.
    b. A class where the ages range between 19 and 50.

The next chapter introduces a measure that reflects how variable the scores are in a distribution by using the deviation scores.

Chapter 2 Practice Test

1. When is the mode the best measure of central tendency?
   a. When the data are measured on an ordinal scale
   b. When the data are discrete
   c. When the data are measure on a nominal scale
   d. When the data are continuous

2. When is the median the best measure of central tendency?
   a. When the data are measured on an ordinal scale
   b. When the data are discrete
   c. When the data are measure on a nominal scale
   d. When the data are continuous

3. When the data are measured on an interval/ratio scale and the scores are normally distributed, what is the best measure of central tendency?
   a. Mean
   b. Median
   c. Mode

4. When the data are measured on an interval/ratio scale and the scores are positively skewed, what is the best measure of central tendency?
   a. Mean
   b. Median
   c. Mode

5. What measure of central tendency is always at the 50th percentile?
   a. Mean
   b. Median
   c. Mode

6. After completing a documentary film about the history of different font styles, the producer wonders if people will find the topic a bit boring. To measure boredom, she asks people to watch the film and then videotapes them while they are watching the film. Later, a research assistant watches the videos and records the number of times each person yawns. The data are at right.

   Compute the mean.
   a. 5
   b. 5.45
   c. 4.5
   d. 6.45
   e. 6
   f. 7

X	f
0	1
1	2
2	1
3	1
4	2
5	6
6	7
7	6
8	5

7. What is the mode for the data described in Question 6?
   a. 5
   b. 5.45
   c. 4.5
   d. 6.45
   e. 6
   f. 7

8. What is the median for the data described in Question 6?
   a. 5
   b. 5.45
   c. 4.5
   d. 6.45
   e. 6
   f. 7

9. What is the sum of the deviation scores for the set of data described in Question 6?
   a. 0
   b. 1
   c. 169
   d. 5.45
   e. It is impossible to answer this question without more information.

10. Which of the following best describes the shape of the distribution of scores described in Question 6? (It may help to sketch the graph.)
    a. Positively skewed
    b. Negatively skewed
    c. Normally distributed

11. What is the best measure of central tendency for the data described in Question 6?
    a. Mean
    b. Median
    c. Mode
    d. All measures are equally appropriate for these data.

12. How do you compute a deviation score?
    a. Subtract the mean from the score.
    b. Add the mean to the score.
    c. Subtract the score from the mean.

13. Which of the following statements about deviation scores is true?
    a. The deviation scores for any given set of data will always add up to zero.
    b. Each individual deviation will always be equal to zero.
    c. The deviation scores will never be negative.

14. A deviation score of 2 indicates that
    a. the X (score) is 2 above the mean.
    b. the mean 2 above the X (score).
    c. the mean will equal 2.

15. Compute the mean for this set of scores: 8, 1, 6, 3, 0, 4, 5, 6, 2
    a. 4.22
    b. 3.89
    c. 4
    d. 6
    e. 5

16. Find the mode for this set of scores: 8, 1, 6, 3, 0, 4, 5, 6, 2
    a. 4.22
    b. 3.89
    c. 4
    d. 6
    e. 5

17. Find the median for this set of scores: 8, 1, 6, 3, 0, 4, 5, 6, 2
    a. 4.22
    b. 3.89
    c. 4
    d. 6
    e. 5

CHAPTER 3

Variability

LEARNING OBJECTIVES

After reading this chapter, you should be able to do the following:

- Explain what the standard deviation measures
- Compute the variance and the standard deviation for a population using a calculator
- Compute the variance and the standard deviation for a sample using a calculator and SPSS

POPULATION VARIABILITY

You have already learned that the mean is commonly used to summarize the center of a distribution of scores measured on an interval or ratio scale. While the mean does a good job describing the center of scores, it is also important to describe how "spread out from center" scores are. For example, imagine you are a psychologist studying individuals' dispositional mood. Some people might have very little change in their mood from day to day, while others might have dramatically different moods on different days. In this case, the *variability* in someone's mood can be really informative to researchers. To illustrate this point, consider two people who both complete the same "happiness scale" for 7 days in a row. A score of 0 = no happiness and a score of 15 = a lot of happiness. George's daily happiness scores are 3, 3, 3, 3, 3, 3, and 3 for Monday through Sunday. In contrast, Morgan's daily happiness scores are more variable. Her Monday through Sunday scores are 2, 2, 5, 2, 6, 4, and 0. Even though the centers of these two data sets are identical (i.e., $\mu = 3$ for both), you would certainly want to describe the different ways that George and Morgan experience mood. Thus, you need to describe the variability, or *"spread,"* of each person's daily mood ratings. There are a number of ways to describe the variability of interval/ratio data. The easiest measure of variability is the range, which is *the difference between the highest and lowest scores*. For example, Morgan's range is $6 - 0 = 6$. The range is a poor measure of variability because it is very insensitive. By insensitive, we mean the range is unaffected by changes to any of the middle scores. As long as the highest score (i.e., 6) and the lowest score (i.e., 0) do not change, the range does not change. A sensitive measure of variability changes if any number in the distribution changes. Researchers value this sensitivity because it allows them to describe the variability in their data more precisely. The most common measure of variability is the standard deviation. *The standard deviation tells you the typical, or standard, distance each score is from the mean.* Therefore, the standard

deviation of George's daily moods is 0 because all of the scores are exactly equal to the mean. In other words, George's daily moods have zero variability. Morgan's daily moods do vary, and therefore, the standard deviation of her data is larger (i.e., 1.93; you will learn to compute this next). Although the standard deviation is the preferred method of measuring variability, it can only be used when the data are interval/ratio. When the data are ordinal, you must use the range.

Reading Question

1. Why is the range a poor measure of variability?
   a. It uses only two values rather than all of the values in the distribution.
   b. It is overly sensitive to changes in the middle of the data.

Reading Question

2. What characteristic of a distribution of scores does a standard deviation describe?
   a. How far scores are from the mean
   b. How spread out the scores are
   c. The variability of scores in a distribution
   d. All of the above

Reading Question

3. The smallest standard deviation that is possible is ____ because this would mean that _____.
   a. −1; all of the scores are negative
   b. 0; all of the scores are the same
   c. 1; all of the scores are positive

Reading Question

4. What measure of variability should be used when the data are ordinal?
   a. Standard deviation
   b. Range

STEPS IN COMPUTING A POPULATION'S STANDARD DEVIATION

We are going to use Morgan's moods to illustrate how to compute the standard deviation. Morgan's daily moods are 2, 2, 5, 2, 6, 4, and 0. We are going to consider these seven scores to be a population because we are only interested in describing this one week of moods. Computing the standard deviation of this population consists of five steps. Focus on understanding what you are trying to do at each step rather than simply doing the calculations.

Step 1: Compute the Deviation Scores ($X - \mu$)

The standard deviation measures the standard (or typical) distance each score is from the mean. Thus, to compute the standard deviation, you first need to determine how far each score is from the mean. The distance each score is from the mean is called a *deviation score* and is computed as $X - \mu$, where X is the score and μ is the mean of the population. For example, this small population of seven scores (2, 2, 5, 2, 6, 4, 0) has a mean of $\mu = 3$. Table 3.1 displays a deviation score for each of the seven scores in the population.

Reading Question

5. A deviation score measures
   a. the typical distance all of the scores are from the mean.
   b. the distance of an individual score from the mean.

Table 3.1 Computing Deviation Scores, Step 1

Score (X)	*Step 1: Deviation Score (X − μ)*
2	2 − 3 = −1
2	2 − 3 = −1
5	5 − 3 = 2
2	2 − 3 = −1
6	6 − 3 = 3
4	4 − 3 = 1
0	0 − 3 = −3

Step 2: Square the Deviation Scores $(X - \mu)^2$

One logical way to find the typical deviation of scores from a mean is finding the average deviation score of a distribution. One could sum the deviation scores and divide their sum by the number of deviation scores, in this case 7. However, if you sum the deviation scores *of any distribution*, you get 0. Of course, if summing deviation scores always yields zero, this approach doesn't help us differentiate between distributions with different amounts of variability. So we need some way to combine deviation scores without losing the variability among the scores. There are a number of ways to avoid this problem, but the one that statisticians use when computing the standard deviation is to square the deviation scores first and then to sum the squared deviation scores.[1] The deviation scores have been squared in Table 3.2.

Table 3.2 Computing Deviation Scores, Step 2

Score (X)	*Step 1: Deviation Score (X − μ)*	*Step 2: Squared Deviation Score $(X - \mu)^2$*
2	2 − 3 = −1	1
2	2 − 3 = −1	1
5	5 − 3 = 2	4
2	2 − 3 = −1	1
6	6 − 3 = 3	9
4	4 − 3 = 1	1
0	0 − 3 = −3	9

1 It is tempting to talk about the standard deviation as the average deviation from the mean, but this is not technically correct because the deviation scores always sum to zero and so the average deviation is 0. A different measure of variability is computed by taking the absolute value of the difference scores. This measure of variability is called the mean absolute deviation. However, the mean absolute deviation is rarely used. You will sometimes hear people talk about the standard deviation as the average deviation. Although this isn't technically accurate, thinking about the standard deviation as the average deviation is fine.

Reading Question

6. Statisticians square each derivation score so that
   a. when they sum them they will not sum to zero.
   b. the standard deviation will be larger.

Step 3: Compute the Sum of the Squared Deviation Scores, $SS = \sum(X - \mu)^2$

Our goal is to compute the typical deviation score of the distribution of scores. Our next step is to compute the **sum of the squared deviation scores (*SS*)**. To compute the *SS*, you simply add (i.e., sum) the squared deviation scores as was done in Table 3.3.

Table 3.3 Computing SS With the Definitional Method, Step 3

Score (X)	*Step 1: Deviation Score* $(X - \mu)$	*Step 2: Squared Deviation Score* $(X - \mu)^2$
2	$2 - 3 = -1$	1
2	$2 - 3 = -1$	1
5	$5 - 3 = 2$	4
2	$2 - 3 = -1$	1
6	$6 - 3 = 3$	9
4	$4 - 3 = 1$	1
0	$0 - 3 = -3$	9
		$SS = \sum(X - \mu)^2 = 1 + 1 + 4 + 1 + 9 + 1 + 9 = 26$

Table 3.3 illustrates how to compute the *SS* using what is called the **definitional formula**, $\sum(X - \mu)^2$. There is another way to find the *SS* that, most of the time, is a lot easier. The second method uses the **computational formula**, $\sum X^2 - \frac{(\sum X)^2}{N}$. Rather than individually computing every score's deviation from the mean, squaring them all, and then summing them all, as the definitional formula requires, the computational formula allows you to find the *SS* with less arithmetic. The computational formula requires you to sum all of the original scores (i.e., the *X*s) to find $\sum X$, square every *X*, and then sum them all to find $\sum X^2$. With this method, you don't need to find each score's deviation from the mean. The computations for this method are shown in Table 3.4.

The computational method and the definitional method will ALWAYS give you the same answer. However, we highly recommend the computational method. Once you get the hang of it, it is much faster. Furthermore, when the mean of the scores is not a whole number (e.g., 3.4578), the definitional formula not only is very tedious but also will lead to rounding error. So, you should work to become proficient with the computational method for finding the *SS*.

Reading Question

7. *SS* stands for the
   a. standard deviation.
   b. sum of the squared deviation scores.
   c. sum of the deviation scores.

Reading Question

8. The definitional method for finding the *SS* and the computational method for finding the *SS* will always provide the same value, but in most situations the _____ method is faster and will not produce rounding error.
   a. Definitional method, $SS = \sum(X - \mu)^2$
   b. Computational method, $SS = \sum X^2 - \frac{(\sum X)^2}{N}$

Table 3.4 Computing *SS* With the Computational Method

Score (X)	*Square Scores* X^2
2	4
2	4
5	25
2	4
6	36
4	16
0	0
$\sum X = 21$	$\sum X^2 = 89$ $SS = \sum X^2 - \frac{(\sum X)^2}{N}$
	$SS = 89 - \frac{(21)^2}{7}$
	$SS = 89 - 63 = 26$

Step 4: Compute the Variance (σ^2)

Again, our goal is to compute the typical, or standard, deviation of the scores from the mean in a distribution of scores. We cannot compute the average deviation score because their sum is always zero. So, instead, we compute the average squared deviation score, which is called the **variance** (σ^2, lowercase sigma squared). When computing any mean, we divide the sum of values by the number of values. Therefore, in this case, we divide the sum of the squared deviation scores by the number of squared deviations (i.e., *N*). The result is the mean of the squared deviation scores, the variance.

$$\text{Population variance: } \sigma^2 = \frac{SS}{N} = \frac{26}{7} = 3.71.$$

Reading Question

9. The variance (σ^2) is the
   a. typical squared deviation from the mean.
   b. typical deviation from the mean.

Step 5: Compute the Standard Deviation (σ)

We squared the deviation scores before we summed them and then divided the sum by *N* to get the variance. This means that the variance is the typical *squared* deviation of all the scores from the mean. While informative, the typical *squared* deviation from the mean is not very intuitive to think about. It is much easier to think about the typical deviation of scores from the mean. Therefore, we convert the

typical *squared* deviation into the typical deviation by taking the square root of the variance. The square root of the variance is the typical or standard deviation of scores from the mean:

$$\text{Population standard deviation: } \sigma = \sqrt{\sigma^2} = \sqrt{\frac{SS}{N}} = \sqrt{\frac{26}{7}} = \sqrt{3.71} = 1.93.$$

The standard deviation tells us the standard (or typical) distance of all the scores from the mean. In this population, the typical distance of all the scores from the mean is 1.93. Some scores are more than 1.93 away from the mean and other scores are less than 1.93 away from the mean, but the "typical" distance of all the scores is 1.93.

Reading Question

10. The standard deviation (σ) is
    a. how far all of the scores are from the mean.
    b. the *typical* distance of all the scores from the mean; some scores will be further away and some closer, but this is the typical distance.

The five steps to computing the standard deviation of a population are listed in Table 3.4. It is worth familiarizing yourself with the verbal labels as well as their symbolic equivalents because we will be using both in future chapters. You should notice that there are two *SS* formulas. While these formulas are mathematically equivalent (meaning they yield the same answer), researchers use the second formula when working with larger data sets. This computational formula is much easier to use with large data sets than is the first definitional formula. You will use both of these equations in a future activity.

Table 3.4 Summary of Five Steps to Computing a Population's Standard Deviation

Population Standard Deviation			
Step	*Verbal Label*	*Symbolic Equivalent*	*Equation*
1	Deviation score		$(X - \mu)$
2	Square the deviation scores		$(X - \mu)^2$
3	Sum of squared deviation scores	SS	Definitional: $SS = \Sigma(X - \mu)^2$ Computational: $SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$
4	Population variance	σ^2	$\sigma^2 = \frac{SS}{N}$
5	Population standard deviation	σ	$\sigma = \sqrt{\frac{SS}{N}}$

Reading Question

11. What symbol represents the standard deviation of a population?
    a. SS
    b. σ
    c. σ^2

Reading Question

12. Which equation defines the sum of the squared deviation scores?
    a. $\Sigma(X-\mu)^2$
    b. $\Sigma(X-\mu)$
    c. $\sqrt{\sigma^2}$

Reading Question

13. Which equation is used for computing the SS?
    a. $SS=\Sigma(X-\mu)^2$
    b. $SS=\Sigma X^2-\frac{(\Sigma X)^2}{N}$
    c. $\sigma=\sqrt{\frac{SS}{N}}$

Once you have computed the standard deviation, you should interpret it in the context of the data set. For this population of Morgan's daily moods, the happiness scores varied. In other words, Morgan was not equally happy every day. The standard deviation indicates how much her happiness varied across the week. Specifically, the standard deviation of 1.93 means that the typical distance of all the happiness scores from the mean of 3 was 1.93. With a mean of only 3, a standard deviation of 1.93 suggests that Morgan's happiness scores varied quite a bit (e.g., 2, 2, 5, 2, 6, 4, 0).

It may help you understand that the standard deviation is actually measuring *the typical distance of all the scores from the mean* if we very briefly consider a completely new data set. Suppose Elliot's average daily happiness score is 9. Specifically, his happiness scores on Monday through Sunday are 8, 8, 11, 8, 12, 10, and 6. Even though he is much happier than Morgan is on a daily basis, the standard deviation of Elliot's daily moods is also 1.93. The standard deviations of these two data sets are identical because both data sets vary equally around their respective means of 3 and 9. Use the space in Table 3.5 to compute the standard deviation of the new data to confirm that it is 1.93.

Table 3.5 Example Table for Computing a Population Standard Deviation

Score (X)	*Squared Score (X^2)*
8	
8	
11	
8	
12	
10	
6	

$$\text{Sum of squared deviations: } SS = \sum X^2 - \frac{(\sum X)^2}{N} =$$

$$\text{Population variance: } \sigma^2 = \frac{SS}{N} =$$

$$\text{Population standard devation: } \sigma = \sqrt{\frac{SS}{N}} =$$

Reading Question

14. In order for two data sets to have the same standard deviation, they must have the same mean.

a. True
b. False

SAMPLE VARIABILITY

Computing the variability of a sample of scores is very similar to computing the variability of a population of scores. In fact, there is only one computational difference that arises when you compute the variance (i.e., Step 4). To highlight the difference between the sample and population formulas, we will analyze the same scores we analyzed earlier (i.e., 2, 2, 5, 2, 6, 4, 0) as if they came from a sample rather than a population.

In the preceding example, we used Morgan's daily moods on each day of a week as if it were a population because we were only trying to describe the variability of Morgan's mood for that week. We were doing descriptive statistics because we were working with data from an entire population. If we wanted to describe the variability of Morgan's moods during all of last year, but she did *not* complete the happiness scale for the entire year, we could use the week's data we have as a sample to estimate the standard deviation of her moods for last year. In this scenario, the week of data we have are a sample from Morgan's entire population of daily moods from last year. In this new scenario, we would be doing inferential statistics, and therefore, there is one small change to how we compute the standard deviation. The reason for the change is that we are using a sample to infer or estimate the value of the population's standard deviation, and the change helps correct for sampling error.

Reading Question

15. When you are using a sample to estimate a population's standard deviation, you are doing _____ statistics.

a. descriptive
b. inferential

Reading Question

16. When computing a sample's standard deviation, there _____ to the computation process relative to when you are computing a population's standard deviation.

a. are many changes
b. is one change

Steps 1 Through 3: Obtaining the *SS*

Computing the sum of the squared deviation scores (*SS*) is identical for a sample and population. The *X* scores were 2, 2, 5, 2, 6, 4, 0. Therefore, $\Sigma X = 21$ and $\Sigma X^2 = 89$.

$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

$$SS = 89 - \frac{(21)^2}{7}$$

$$SS = 26.$$

Reading Question

17. The *SS* is computed in exactly the same way for a sample and a population.
    a. True
    b. False

Step 4: Compute the Sample Variance (SD^2)

Although the *SS* computations are the same, there is a difference between the computation of a sample variance and a population variance. To compute the *population variance,* you divided the *SS* by *N*. To compute the *sample variance,* you divide the *SS* by $N - 1$. This is the only difference between the computation of a variance for a sample and a population:

$$\text{Sample variance: } SD^2 = \frac{SS}{N-1} = \frac{26}{6} = 4.33.$$

Why we divide by $N - 1$ when using a sample to estimate a population's variability rather than by *N* is a somewhat complicated issue. The simplest explanation is that samples are less variable than populations, and without the $N - 1$ adjustment, our variability estimate would be too low. For example, the variability of Morgan's daily moods during a 7-day period, our sample, will be less than her moods during a 365-day period, our population. The small 7-day sample is going to have less variability than will the much larger 365-day population. More data tend to create more variability. The difference in variability between smaller samples and larger populations is a serious problem if you are trying to use a sample to estimate a population's standard deviation. So, you need to do some kind of computational adjustment when using a sample to estimate a population's variability; if you don't, your variability estimate will tend to be too low. The computational adjustment statisticians determined to be most accurate in most situations is dividing the *SS* by $N - 1$ rather than by *N*.

Reading Question

18. When using a sample to estimate a population's variability, the *SS* is divided by $N - 1$ rather than by *N* to correct for a sample's tendency to
    a. overestimate the variability of a population.
    b. underestimate the variability of a population.

Step 5: Compute the Sample Standard Deviation (*SD*)

Take the square root of the sample variance:

$$\text{Sample standard deviation: } SD = \sqrt{\frac{SS}{N-1}} = \sqrt{\frac{26}{6}} = \sqrt{4.33} = 2.08.$$

The verbal labels corresponding to each computational step for a sample's standard deviation are identical to those used when computing a population's standard deviation. However, as indicated above, the sample's symbolic equivalents are Arabic letters rather than Greek letters (Table 3.6).

Table 3.6 Summary of Five Steps to Computing a Sample's Standard Deviation

Sample Variability			
Step	*Verbal Label*	*Symbolic Equivalent*	*Equation*
1	Deviation score		$(X - M)$
2	Square the deviation scores		$(X - M)^2$
3	Sum of squared deviation scores	*SS*	Definitional: $SS = \Sigma(X - M)^2$ Computational: $SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$
4	Sample variance	SD^2	$SD^2 = \frac{SS}{N-1}$
5	Sample standard deviation	*SD*	$SD = \sqrt{\frac{SS}{N-1}}$

Reading Question

19. What symbol represents the standard deviation of a sample?
   a. *SD*
   b. SD^2
   c. *SS*

Reading Question

20. When *computing* the variance of an entire population, you are performing _____ so divide the *SS* by ___.
   a. Descriptive statistics, *N*
   b. Inferential statistics, $N - 1$

Reading Question

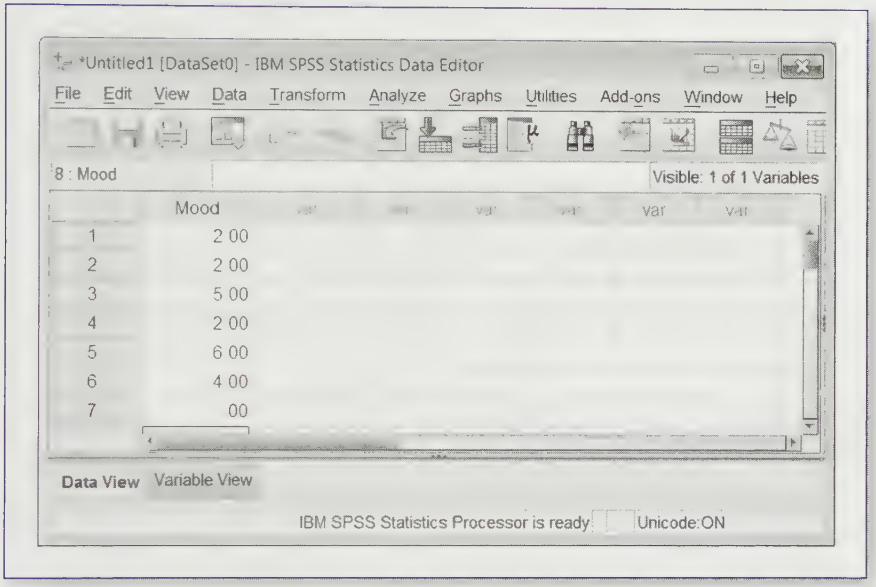
21. When *estimating* the variance of a population from a sample, you are performing _____ so divide the *SS* by ___.
   a. Descriptive statistics, *N*
   b. Inferential statistics, $N - 1$

SPSS

Data File

You can compute the standard deviation and variance of a *sample* (not a population) using SPSS. Begin by entering the sample scores (2, 2, 5, 2, 6, 4, 0) into one column in SPSS. This is what your data file should look like when it is done (Figure 3.1):

Figure 3.1 SPSS Screenshot of Data Entry Screen



Computing Measures of Variability

- Click on the Analyze menu. Choose Descriptive Statistics and then Frequencies (see Figure 3.2).
    - You can obtain descriptive statistics (e.g., mean, standard deviation) in a lot of different ways in SPSS. We are only showing you one way here, but if you explore the menus, you can find other ways to obtain the same statistics.
- Move the variable(s) of interest into the Variable(s) box (see Figure 3.3).
- Make sure the Display Frequency Tables box is unchecked if you do not want a frequency distribution table.
- Click on the Statistics button.
- Click on the boxes for mean, standard deviation, variance, minimum, and maximum, and then click on the Continue button, and then click OK (see Figure 3.4).
- **Important Note: SPSS** computes the sample standard deviation and variance, not the population values.

Figure 3.2 SPSS Screenshot of Analyze Menu for Measures of Variability

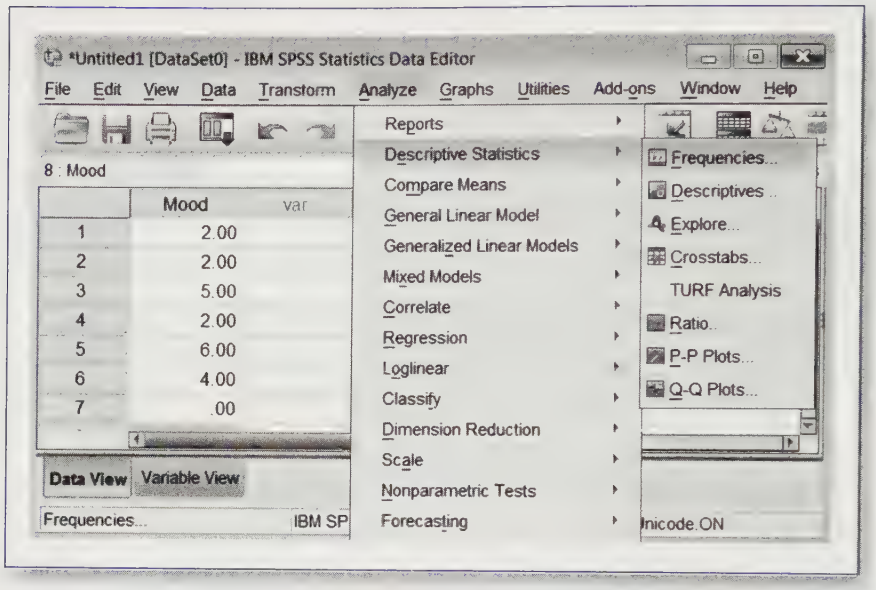


Figure 3.3 SPSS Screenshot of Choosing Variables for Analysis

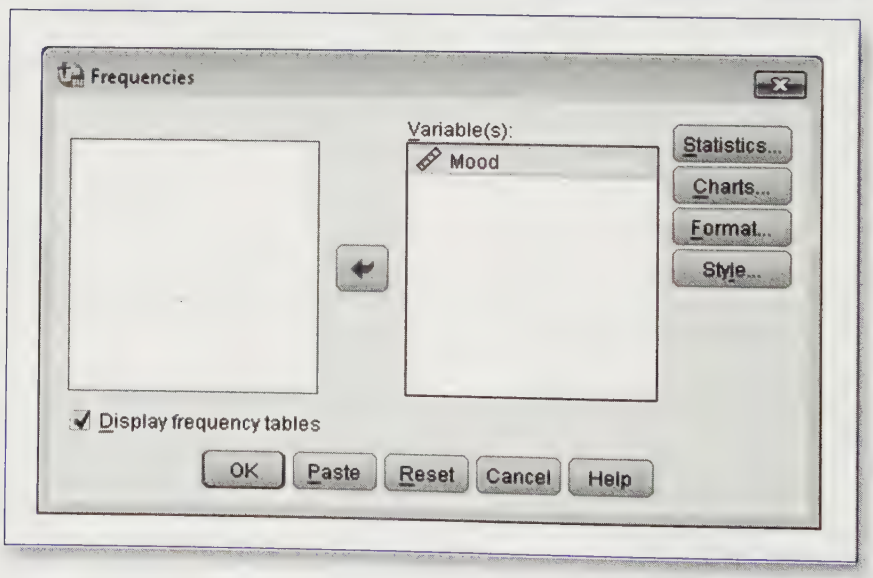
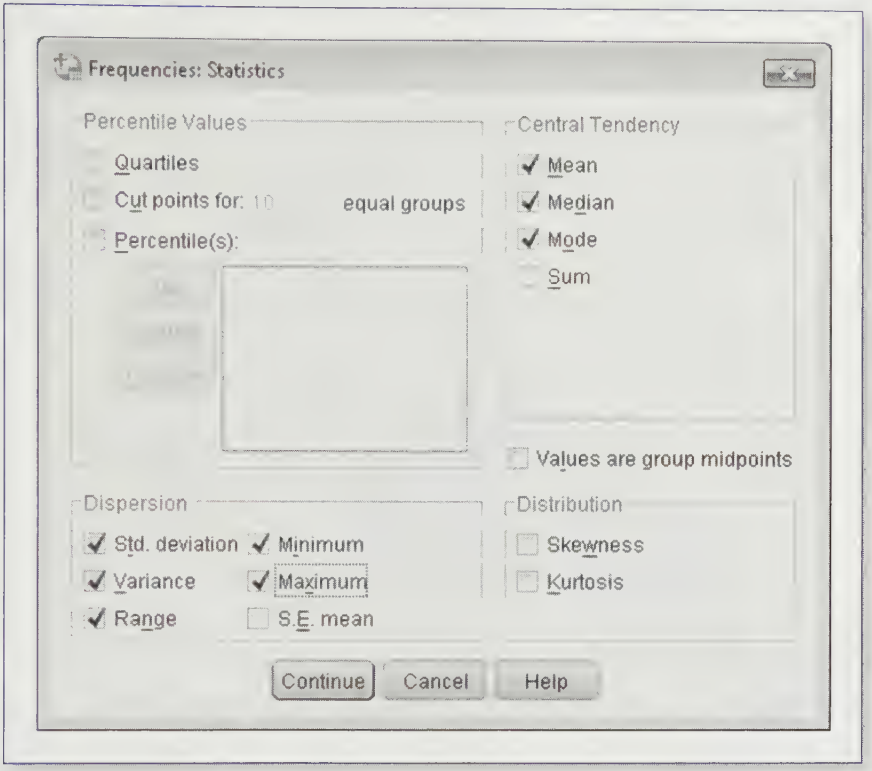


Figure 3.4 SPSS Screenshot of Selecting Descriptive Statistics



Output

Your output file should look similar to the one below. Note that the results are the same as what you did by hand (Figure 3.5).

Reading Question

22. What is the standard deviation of these data?
    a. 3
    b. 2
    c. 2.08

Reading Question

23. SPSS can only be used to compute the standard deviation of a sample, not a population.
    a. True
    b. False

Figure 3.5 SPSS Output for Descriptive Statistics

Statistics

Mood

N	Valid	7
	Missing	0
Mean		3.0000
Median		2.0000
Mode		2.00
Std. Deviation		2.08167
Variance		4.333
Minimum		.00
Maximum		6.00

OVERVIEW OF THE ACTIVITY

In Activity 3.1, you will work with data sets to better understand variability and what causes variability in a data set. You will also practice computing the *SS* and the standard deviation for populations and samples using your calculator.

Activity 3.1: Variability

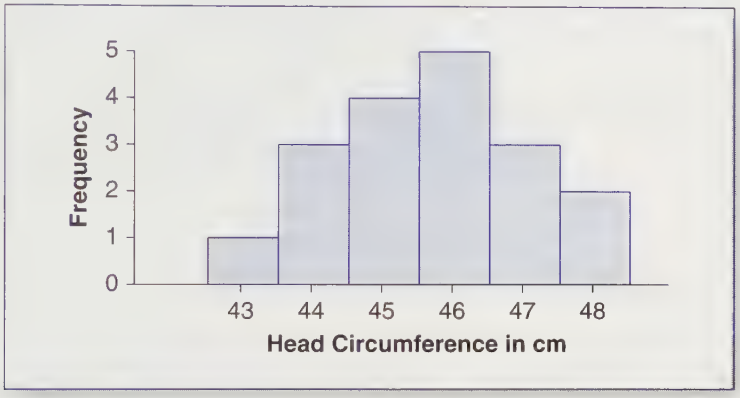
Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Recognize how measurement error, individual differences, and treatments can create variability
- Determine which of two distributions has a higher standard deviation by comparing histograms
- Explain what the standard deviation is to someone who has not taken a statistics course
- Compute the standard deviation for population or sample data presented in a histogram or a frequency table
- Use the definitional and computational formulas to compute the *SS*
- Use the statistics mode on your calculator to find the $\sum X^2$ and the $\sum X$

PART I: CONCEPTUAL UNDERSTANDING OF VARIABILITY

A developmental psychologist studying how the neonatal environment of infants affects their development is planning a huge national study. He needs to develop a reliable way to collect the physical measurements of newborn infants. He knows that the physical development of infants is carefully tracked during the child's first year of life. At every doctor's visit, the child's height, weight, and head circumference are measured, typically by nurses. Consequently, he approaches nurses because he hopes that he can use the data they will collect from their future patients in his study. However, before he does, he wants to know how accurate their measurement procedures actually are. After getting the nurses' agreement to participate, he brings a very realistic doll of a 1-year-old infant to the nurses and had each nurse measure the circumference of the infant's head in centimeters. The 18 nurses' measurements are graphed below.



1. Given that all nurses were measuring the head of the same doll, all of the head circumference measurements should be the same. In other words, there should be no variability in the measurements. However, the above graph clearly illustrates that there was variability in the measurements. Why was there variability in the measurements of the doll's head circumference? (Select all that apply.)
   a. Some nurses held the tape measure tighter around the infants head while others held it looser.
   b. Each nurse put the tape measure in a slightly different place on the doll's head
   c. Some nurses may have misread the tape measure.
   d. The doll's head changed size between measurements.

2. In the above question, all of the variability in scores was created by **measurement error** because everyone was measuring the same thing and, therefore, should have obtained the same score. Unfortunately, measurement error is always present. No matter what you are measuring, you will never be able to measure it perfectly every time. You can, however, reduce the amount of measurement error. In the context of measuring an infant's head circumference, how could the developmental psychologist and/or nurses reduce the variability in scores created by measurement error (i.e., what could they do to increase the accuracy/reliability of each measurement?). Select all that apply.

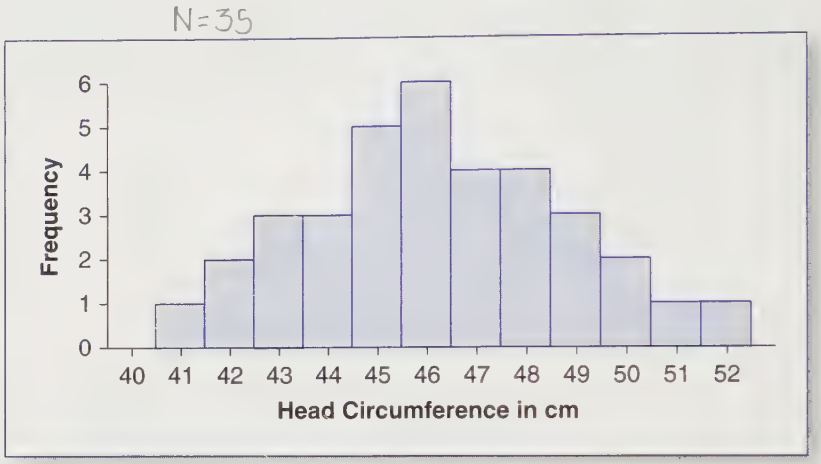
(a.) Give the nurses a lot of practice measuring different dolls' heads.

(b.) Train the nurses to use a consistent degree of tension in the tape measure.

c. Use dolls with heads made out of a soft, pliable material.

(d.) Use a tape measure that only records centimeters, not inches.

The following week, different nurses measured the head circumference of 35 different infants. The head circumference measurements for 35 *different* infants are graphed below:



3. Did measurement error create some of the variability in scores that are graphed above?

(a.) No, there is no measurement error. The head circumferences are only different because different infants were measured.

b. Yes, there is measurement error. There is always some potential for measurement error any time a measurement is taken.

4. In addition to measurement error, something else is also creating variability in the distribution of 35 scores (i.e., head circumferences). Besides measurement error, what is another reason for the variability in the above distribution of 35 infants' head circumferences? (Select all that apply)

a. Different nurses measured the head circumferences and each nurse may have used a slightly different measurement technique.

(b.) The 35 infants have heads that vary in size.

In the previous question, the variability in head circumferences was created by both *measurement error,* which is always present, and the fact that the 35 infants' heads actually varied in size. Researchers refer to this second source of variability as being created by **individual differences.** The fact that people are different from each other creates variability in their scores.

5. Using highly standardized measurement procedures can reduce the amount of variability created by _____.

a. individual differences

(b.) measurement error

6. In which of the following distributions of scores would there be *more* variability created by individual differences?
   a. The heights of 50 first graders
   b. The heights of 50 elementary school children (first through fifth graders)

It should be clear to you that researchers need to understand the variability in their data and what is creating it. Researchers view measurement error variability as "bad" and attempt to *minimize* it. Researchers also recognize that individual differences variability will always create variability in their data, and they try to control this variability with carefully designed experiments. In addition, in many research situations, researchers actually want to *generate* variability by creating different kinds of treatments.

For example, suppose the researcher thought that physically touching prematurely born infants would increase their growth. To test this hypothesis, the researcher could conduct a study with two samples of prematurely born infants. All of the infants in Group 1 could be touched with skin-to-skin contact for at least 6 hours a day. All of the infants in Group 2 could be touched only by someone wearing gloves. After 4 weeks of these differing **treatments,** the circumferences of the babies' heads could be compared.

7. In this study, there are three things creating variability in infants' head circumference. The fact that measuring an infant's head circumference is hard to do accurately contributes to the amount of _____ in this study.
   a. treatment differences variability
   b. individual differences variability
   c. measurement error variability

8. The fact that the researcher gave some infants 6 hours of skin-to-skin touch a day and some other infants no skin-to-skin touch contributes to the amount of _____ in this study.
   a. treatment differences variability
   b. individual differences variability
   c. measurement error variability

9. The fact that infants naturally differ from each other in head size contributes to the amount of _____ in this study.
   a. treatment differences variability
   b. individual differences variability
   c. measurement error variability

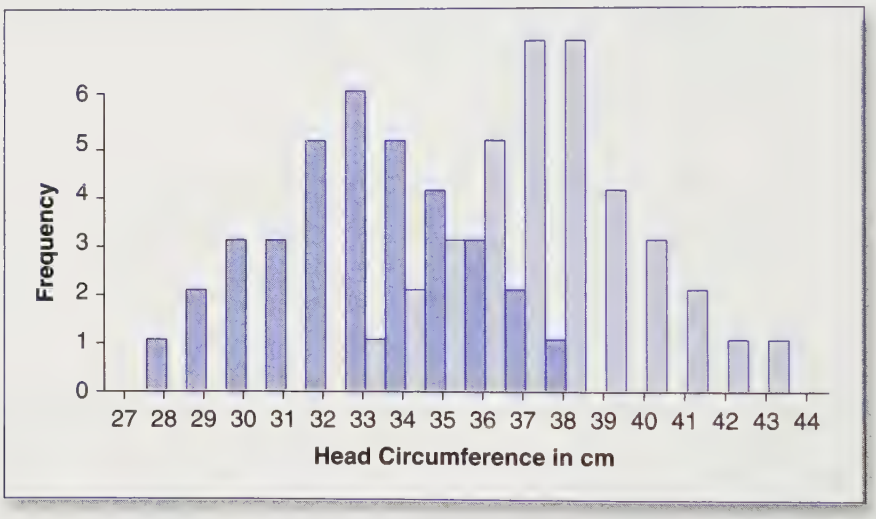
10. If we measured each of the following variables for every person in this class, which variables would have the most *measurement error variability?*
    a. Students' report of their parents' annual income
    b. Parents' annual income recorded from official tax forms

11. If we measured each of the following variables for every person in this class, which variables would have the least *individual differences variability?*
    a. Number of siblings a person has
    b. Number of fingers a person has

12. Understanding variability is important because some variables simply have more variability than others do. For example, in high school students, which of the following variables would have the largest standard deviation?
    a. Annual income of parents
    b. Age

13. Which of the following variables would have the smallest standard deviation for high school students?
    a. Number of phone calls made in a day
    b. Number of phones owned

The following figure displays the head circumferences of 70 premature infants. Half of the infants were only touched by someone wearing gloves (*the darker bars*). The other half of the infants were only touched by someone who was not wearing gloves (*the lighter bars*).



14. In the above figure, the variability created by the different treatments (i.e., touching infants while wearing gloves vs. touching infants while not wearing gloves) is depicted by the fact that
    a. all of the infants who were touched while wearing gloves do not have the same head circumference.
    b. all of the infants who were touched while not wearing gloves do not have the same head circumference.
    c. the infants who were touched without wearing gloves (*lighter bars*) tended to have larger head circumferences than infants who were touched while wearing gloves (*darker bars*).

15. In most research situations, there will be variability that is created by *measurement error, individual differences,* and *differing treatments*. In the study described earlier, the researcher expected that touch would result in faster growth. Thus, the researcher compares the mean head circumference for a sample of the premature babies who were touched with direct skin contact to the mean head circumference for a sample of the premature babies who were only touched by someone wearing gloves. Suppose that the mean head circumference for the direct touch sample was 38 cm, and the mean for the other sample was 33 cm. Why can't we just look at those two numbers and conclude that direct skin touching facilitated infant growth? Select all that apply.

a. The variability between the sample means may have been created by a treatment effect.

b. The variability between the sample means may have been created by individual differences.

c. The variability between the sample means may have been created by measurement error.

A primary goal of this course is to teach you how researchers determine if the variability you see in data (e.g., the difference between the head circumferences of infants who were touched in different ways) was likely created by a treatment difference or if the variability was likely created by individual differences and/or measurement error differences that exist between treatment conditions (i.e., sampling error).

16. The primary goal of this course is to teach you how to

a. design experiments to test treatment effects.

b. determine if variability is likely to be due to treatment effects or sampling error.

c. eliminate measurement error and individual difference variability.

As you know from the reading on variability, the standard deviation is commonly used to measure the amount of variability in a set of scores. Computing the standard deviation will not enable you to determine if the variability in the scores is created by treatment differences, individual differences, or measurement error. The standard deviation reveals the typical variability of scores from their mean. Later in the course, we will learn how to use other statistics to help us determine if treatment differences created variability in scores.

17. The standard deviation is a measure of

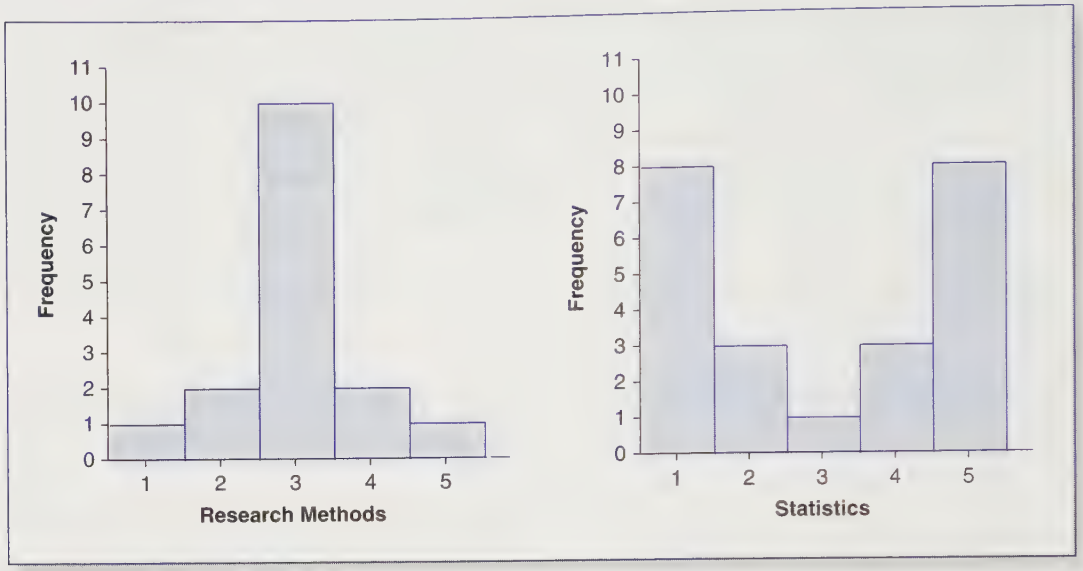
a. treatment variability in scores.

b. individual differences variability in scores.

c. typical distance of scores (i.e., variability) from the mean score.

If you understand the concept of variability, you should be able to "read" histograms. Specifically, you should be able to determine which of two histograms has more variability (i.e., a higher standard deviation).

For example, suppose that a professor asked students at the end of the semester how much they agree with the statement, "I enjoyed taking this course." Students may respond with 1 = *strongly agree,* 2 = *agree,* 3 = *neither agree nor disagree,* 4 = *disagree,* or 5 = *strongly disagree*. Distributions from two of his classes are displayed on page 84. The first graph is from a research methods course, and the second graph is from a statistics course



18. You should note that the mean rating for both courses was 3 (*neither agree nor disagree*). Which of the courses had more scores closer to its mean?

    a. Research methods

    b. Statistics

19. Given that the standard deviation measures the typical distance of scores from the mean, which course has the *smaller* standard deviation? (Hint: Both distributions have a mean of 3.)

    a. Research methods

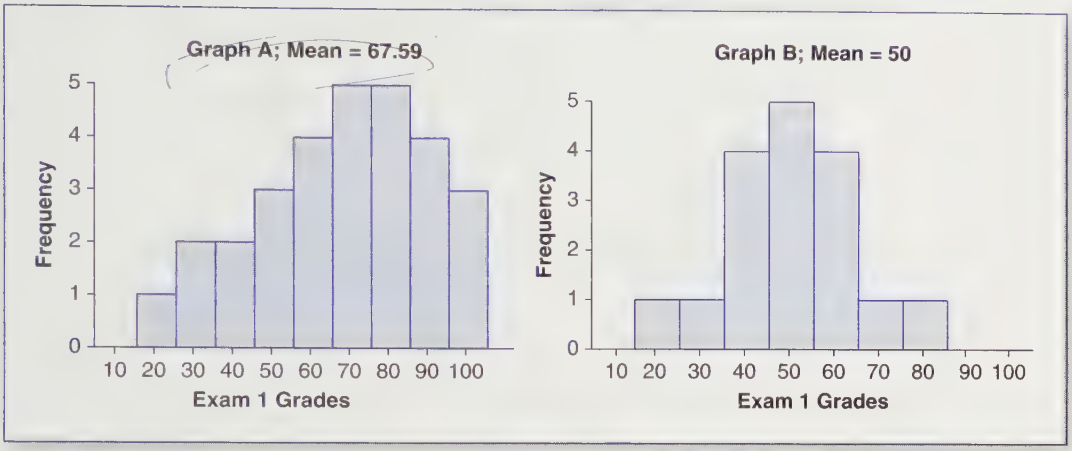
    b. Statistics

There is far less variability in the research methods course than in the statistics course. In the research methods course, the majority of students responded with a 3, and most responses were very close to the mean. However, in the statistics course, most people responded with either a 1 or a 5, and most responses were relatively far from the mean. In other words, in the research methods course, most students gave the same answer (i.e., there was little variability in their responses). However, in the statistics course, there were greater differences of opinion (i.e., there was a lot of variability in responses). In general, *graphs with a lot of data points "piled up" close to the mean (like the research methods distribution) have less variability (i.e., a smaller standard deviation) than graphs with a lot of data points "piled up" further from the mean (like the statistics distribution).*

While there are other factors to consider, looking at where the scores pile up relative to the mean is a good way to start "reading" the variability in a distribution of scores. Use this rule to "read" the variability in the following pairs of graphs.

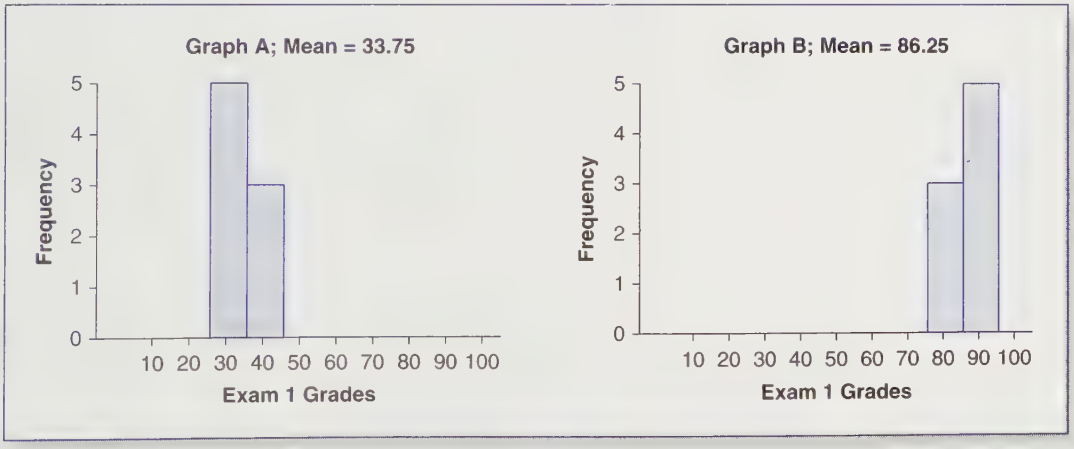
For Questions 20 to 22, determine if Graph A has more variability, Graph B has more variability, or if they have similar amounts of variability.

20.



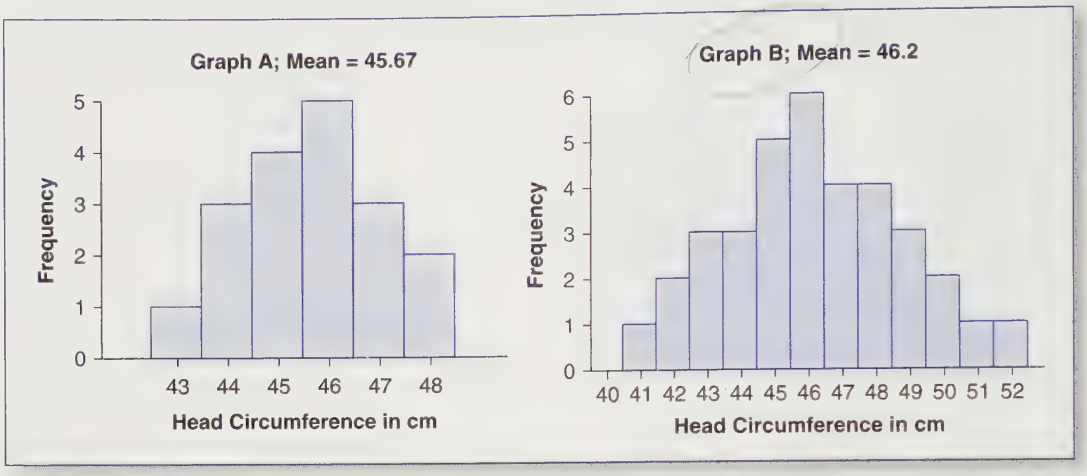
Explain your choice:

21.



Explain your choice:

22.



Explain your choice:

23. If a histogram has many scores piled up close to the mean value, the data set will tend to have
    a. a large standard deviation.
    b. a small standard deviation.

You should also note that there is no cutoff value for large or small standard deviations. In this case, the standard deviation for the research methods course was $SD = 0.89$, and the standard deviation for the statistics course was $SD = 1.78$. We can say that the standard deviation for the statistics class was relatively large because a standard deviation of 1.78 is large when the range of possible responses is only between 1 and 5. A typical distance of 1.78 from the mean on a 5-point scale is quite large. However, if teaching evaluations were made on a 50-point scale, a standard deviation of 1.78 would be quite small.

24. Which of the following standard deviations would represent greater variability relative to the range of possible scores?
    a. A standard deviation of 2.51 for a variable measured with a 1-to-7 Likert scale
    b. A standard deviation of 2.51 for a fifth-grade spelling test (scores could potentially vary between 0 and 100)

PART II: COMPUTING THE STANDARD DEVIATION

A group of four students reports their heights in inches as follows: 68, 61, 72, 70.

25. Use the table below to help you compute the *SS* (sum of the squared deviation scores) using the definitional formula:

M = 67.75
N = 4

$$SS = \Sigma(X - M)^2$$

Score (X)	(X – M)	$(X - M)^2$
68	68 - 67.75 = .25	.06
61	61 - 67.75 = -6.75	45.56
72	72 - 67.75 = 4.25	18.06
70	70 - 67.75 = 2.25	5.06
		$SS = \Sigma(X - M)^2 =$ 68.74

You should find that the M is 67.75 and the SS is 68.75.

26. Although the definitional formula makes intuitive sense, it is not an easy formula to work with when you have a large set of scores. With large data sets, it is far easier to compute the SS using the **computational formula:**

$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

The definitional and computational SS formulas will yield identical values. To avoid a common error when using the computational formula, be sure you understand the distinction between ΣX^2 and $(\Sigma X)^2$. To compute ΣX^2, you should square each score first, then sum them ($X^2 = 68^2 + 61^2 + 72^2 + 70^2$). To compute $(\Sigma X)^2$, you should sum all the scores first and then square the sum ($(\Sigma X)^2 = (68 + 61 + 72 + 70)^2$). The N is the number of scores. Use the table below to help you compute the SS using the computational formula.

Score (X)	X^2
68	4,624
61	3721
72	5,184
70	4900
$\Sigma X =$	$\Sigma X^2 =$ 18,429

$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

$$18429 - \frac{(271)^2}{4} \quad \frac{73,441}{4}$$

$$18429 - 18,360.25$$

68.75

This is a very small data set, so it is probably not obvious that the computational formula for the SS can save you quite a bit of time. When working with large data sets or when the mean of the data is not a whole number, the definitional formula takes longer, and the final answer is likely to have rounding error. Another advantage to using the computational formula is that even cheap statistics calculators will compute the ΣX^2 and ΣX for you. Therefore, if you learn how to use your statistics calculator, computing the SS will become quite easy. You can simply substitute the values of ΣX and ΣX^2 into the computational formula. Try to use the statistics mode on your calculator to find the ΣX^2 and ΣX.

27. Use the *SS* you computed in Question 26 to compute the standard deviation, assuming the data came from a sample. *You will use the following equation whenever you are analyzing data from a sample.* You should get 4.79.

$$SD = \sqrt{\frac{SS}{N-1}}$$

28. Now use the *SS* you computed in Question 26 to compute the standard deviation, assuming that the data came from a population. *You will use the following equation whenever you are analyzing data from an entire population.* You should get 4.15.

$$\sigma = \sqrt{\frac{SS}{N}}$$

29. Figure out how to use the statistics mode on your calculator to compute the standard deviation of a population and a sample. There should be one button you can push or one line in a display that shows you the sample and population standard deviation. Don't skip this. Finding the standard deviation with your calculator will be extremely helpful later in the course.

30. Compute the *SS* and the standard deviation for the following sample of five scores: 5, 6, 3, 2, 7.

Compute the SS using the definitional formula		
Score (X)	*(X – M)*	*$(X - M)^2$*
5	.4	.16
6	1.4	1.96
3	−1.6	2.56
2	−2.6	6.76
7		5.76
		$SS = \Sigma(X - M)^2 =$

Confirm that you obtain the same results using the computational formula:

$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N}.$$

Compute the standard deviation:

$$SD = \sqrt{\frac{SS}{N-1}}.$$

31. In the previous examples, you computed the *SS* using both the definitional and computational formulas. Although the definitional formula makes intuitive sense, it is far easier to use the computational formula. For all subsequent problems, you should use the computational formula. Compute the *SS* and the standard deviation for the following population of five scores: 1, 3, 3, 5, 7.

$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N}.$$

Compute the standard deviation:

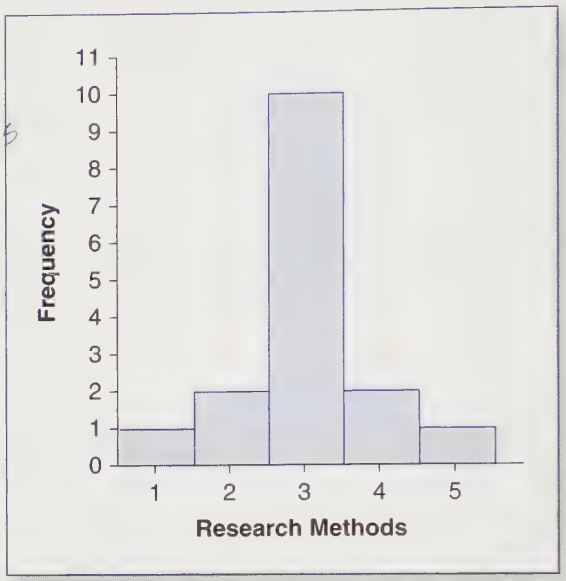
$$\sigma = \sqrt{\frac{SS}{N}}.$$

32. In most situations, which of the following formulas should you use to compute *SS*?

a. $SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$

b. $SS = \Sigma (X - M)^2$

33. The graph from the research methods course described earlier is reproduced below. Create a frequency distribution table from these population data.



34. Compute the standard deviation of the ratings in the research methods course.

35. You should have found that the standard deviation was 0.87. What does 0.87 mean in the context of this set of data?

    a. The typical distance between scores is .87.

    b. The typical distance between the scores and the mean is .87.

    c. The typical distance between the sample means is .87.

36. After computing the standard deviations for the research methods course, the instructor realizes that some students did not complete the rating form, and so it was a sample, not an entire population. Recompute the standard deviation of the course as if it came from a sample.

Chapter 3 Practice Test

1. A teacher asks nine of his students to read a book for 15 minutes, and he records the number of lines of text each student reads during that 15 minutes. The results are as follows:

   20, 13, 33, 11, 40, 29, 15, 38, 21 / 9 = 24.44 N = 9

   What is the mean for this sample of nine students?

   a. 22.00
   b. 24.44
   c. 10.91
   d. 22.68

2. What is the deviation score for the person who read 20 pages?

   a. 24.44
   b. −24.44
   c. −4.44
   d. 4.44

3. What is the standard deviation for this sample of nine students?

   a. 24.44
   b. 10.91
   c. 10.29
   d. 220
   e. 952.22
   f. 14.88

   400 + 169 + 1089 + 121 + 1600 + 841 + 1444 + 441 + 225 = 6330

   $6330 - \frac{48400}{9}$

   $6330 - 5377.777 = \sqrt{\frac{952.22}{8}} = 10.29$

4. Which of the following statements is the best interpretation of a standard deviation?

   a. The typical distance deviation scores are from the population
   b. The typical distance scores are from each other
   c. The typical distance scores are from the deviation scores
   d. The typical distance scores are from the mean

5. What is *SS*?

   a. The sum of the squared deviation scores
   b. The sum of the squared scores
   c. The sum of the squared standard deviations
   d. The square of the summed scores

6. When is the standard deviation not an appropriate measure of variability?

   a. When the data are nominal or ordinal
   b. When the data are interval or ratio
   c. When the data are leptokurtic
   d. When the data are normally distributed

7. Compute the standard deviation for this population of scores:

X	f
100	2
90	5
80	3
70	0
60	1

a. 11.20
b. 86.36
c. 950
d. 10.68

83300 − 902,500 = 10.68

83,200 − 82045 = 1254.55 (45)

100,100,90+90+90+90+90+80+80+80+60 =950

mean = 86.36

10,000+10000+8100+8105+ + + + 6400 + 3600

40,500 + 19200 + 5 = 83,300

8. An instructor gives a multiple-choice exam that is graded electronically. What type of error is this instructor minimizing by using a multiple-choice, electronically graded exam?
   a. Individual differences
   b. Treatment effects
   c. Measurement error
   d. All of the above sources of variability would be reduced.

9. Which of the following variables would have *more* variability due to individual differences?
   a. SAT scores of college students in the nation
   b. SAT scores of college students at your college

10. The scores on the first exam in a statistics course had a mean of 78.32 with a standard deviation of 13.24. Scores on the second exam had a mean of 80 with a standard deviation of 11.32. For which exam were scores closer to the mean?
    a. Exam 1
    b. Exam 2

11. Scores on an exam for students in the same section of a chemistry course had a mean of 91.34 with a standard deviation of 14.63. Which of the following are sources of the variability in the exam scores? You may select more than one.
    a. Individual differences
    b. Treatment effects
    c. Measurement error

12. At the beginning of the school year, all of the students in three third-grade classes take a test to assess their current math skills. The mean and the standard deviation for each class are as follows:

    Class 1: Mean = 73.59, *SD* = 24.78

    Class 2: Mean = 65.42; *SD* = 8.43

    Class 3: Mean = 85.32; *SD* = 22.86

Which class do you think will be easiest to teach?

a. Class 1. The standard deviation is highest in Class 1, indicating that the students in that Class have higher math skills than the students in the other classes.

b. Class 2. The standard deviation is lower for this class than Class 1 or 3. This suggests that the students are more homogeneous (similar) in their math skills and that it will be easier for the teacher to create lessons that will work for most of the students.

c. Class 3. The mean test score was highest for this group of students, which suggests that these students have the highest math skills and the teacher can create lessons that will work well for this group of high-ability students.

13. Can you use the statistics mode on your calculator to compute the mean and standard deviation for a set of data?

a. Yes, no problem!

b. I can do it if I have a set of instructions in front of me.

c. No, I haven't been able to figure it out yet.

CHAPTER 4

z Scores

LEARNING OBJECTIVES

After reading this chapter, you should be able to do the following:

- Compute and interpret a *z* score for a given raw score
- Solve for *X* if given a *z* score
- Explain what the sign of a *z* score indicates
- Explain what the absolute value of a *z* score indicates
- Locate a *z* score within a distribution
- Use a unit normal table to determine the proportion of scores above or below any given *z* score

z FOR A SINGLE SCORE

In Chapter 2, you learned that the mean perfectly balances the positive and negative deviation scores of a distribution. In other words, when you compute each score's distance from the mean, the sum of the positive deviation scores will always equal the sum of the negative deviation scores. In Chapter 3, you learned that the standard deviation describes how much variability there is in a set of numbers. Together, the mean and the standard deviation help you interpret a distribution of scores by telling you the "center" of the scores and how much scores vary around that center.

In a very similar way, the mean and the standard deviation can also help you interpret an individual score in a distribution. For example, suppose your score on the ACT was 25. This score alone doesn't tell you much about your performance, but if you knew that the mean ACT score was 21 with a standard deviation of 4.70, you could interpret your score. Your score of 25 was 4 points better than the population mean. The population standard deviation was 4.70; this means that your score of 25 (i.e., +4 from the mean) deviated less from the mean than was typical (4.70). So you did better than average but only a little better because your score was less than 1 standard deviation above the mean.

Clearly, knowing your score's deviation from the mean, whether that deviation is positive or negative and whether the deviation is smaller or larger than the standard deviation, helps you interpret your ACT score. In this chapter, you will learn to use the *z* for a single score. This statistical procedure is the mathematical equivalent of using the mean and the standard deviation of a distribution to help you

interpret an individual score. As is often the case with mathematical procedures, it has the advantage of being more precise than the logical analysis provided in the previous paragraph.

Reading Question

1. The z for a single score is a procedure that
   a. tells if a score is above or below a population mean.
   b. informs whether a score's deviation from the mean is relatively large or relatively small compared with the deviations of the rest of the data.
   c. both of the above.

The z for a single score is useful for two purposes. First, as described earlier, it is used to locate a score in a distribution of scores. A z score will indicate if a given score is very good (far above the mean), very bad (far below the mean), or average (close to the mean). For example, when looking at ACT scores, larger positive z scores represent better performance and larger negative z scores represent worse performance. In the following section, you will learn how to compute and interpret z scores so you can evaluate an individual's performance. Second, a z for a single score can help you compare two scores from *different* distributions. For example, the z for a single score can help you compare scores on the ACT and the SAT even though the maximum score on the ACT is 36 and the maximum score on the math and verbal sections of the SAT is 1600. The z for a single score "corrects" for the differences in scale. Therefore, z scores can reveal if a score of 22 on the ACT was better than a score of 1100 on the SAT.

Reading Question

2. The z for a single score can be used to
   a. locate a specific score in a distribution of scores and compare scores from different distributions.
   b. locate a sample mean in a distribution of means and compare means from different distributions.

Computing a z for an Individual Score

To compute a z for a single score, you need to know the score (X), the mean of the population (μ), and the standard deviation of the scores. While the z for a single score can be computed with either the population standard deviation (σ) or a sample standard deviation (SD), the sample standard deviation is used only when the population standard deviation is not known. Likewise, the sample mean (M) is used only when the population mean (μ) is not known. For example, suppose that a student took the ACT and obtained a score of 22. The mean for this test is $= 21$, with a standard deviation of $\sigma = 4.70$. To compute the z score, you first compute the deviation between the score and the population mean ($X - \mu$). You then divide this difference by the standard deviation—in this case, the population standard deviation (σ):

$$z = \frac{X - \mu}{\sigma} = \frac{22 - 21}{4.70} = 0.21.$$

Interpreting the z for a Single Score

The z score is positive because the student's score of 22 was above the mean ACT score of 21. The numerator of the z formula is a deviation score ($X - \mu$). It indicates how far the given score is above or

below the population mean. Thus, whenever a raw score is above the mean, you will obtain a positive *z* score, and whenever a raw score is below the mean, you will obtain a negative *z* score.

The denominator of the *z* formula is the population standard deviation (σ). The standard deviation is the typical distance scores are from the mean in the population. Using the population standard deviation as the denominator of the *z* score equation means that *z* scores should be interpreted as the proportion of standard deviation units a given raw score is away from the mean of the distribution.

So, as described earlier, the *z* score reveals two really important bits of information. First, if it is positive, the given score was above the average score. In addition, the farther the absolute value of the *z* score is from 0, the more the score deviates from the mean. A *z* score with an absolute value of 1 indicates that the score is 1 standard deviation from the mean. The standard deviation is the average, or typical, distance of scores from the mean. Thus, if the *z* score is greater than 1 (or less than −1), the score deviates from the population mean more than is typical. In this case, the student's *z* score of +0.21 was above average but only slightly so.

Reading Question

3. If a given *z* score is negative, then the raw score it is representing was
   a. above the population mean.
   b. below the population mean.
   c. at the population mean.

Reading Question

4. Which of the following *z* scores represents a raw score that is the most atypical (i.e., farthest from the mean)?
   a. −3.10
   b. +2.20
   c. −0.81
   d. +0.47

Reading Question

5. A *z* score of +1.90 is above the mean more than is typical.
   a. True
   b. False

Reading Question

6. A *z* score of +1 is above the population mean by exactly 1 standard deviation.
   a. True
   b. False

USING *X* TO FIND IMPORTANT "CUT LINES"

In some situations, you might be interested in identifying the specific score that is well above (or below) the average so you can use that score as a "cut line." For example, perhaps you are interested in identifying students who score "far above average" on the ACT so you can invite them to join an honors program. You might decide that you want to know what score represents performance that is 2 standard deviations above the average score of 21 on the ACT. You could do this by using the *z* formula, inserting 2 for *z* and 21 for μ. You would also have to know the standard deviation (i.e., σ or *SD*).

Here is an example of solving for the raw score when given that $z = 2$, $\mu = 21$, and $\sigma = 4.70$:

$$z = \frac{X - \mu}{\sigma}$$

$$2 = \frac{X - 21}{4.70}$$

$$9.4 = X - 21$$

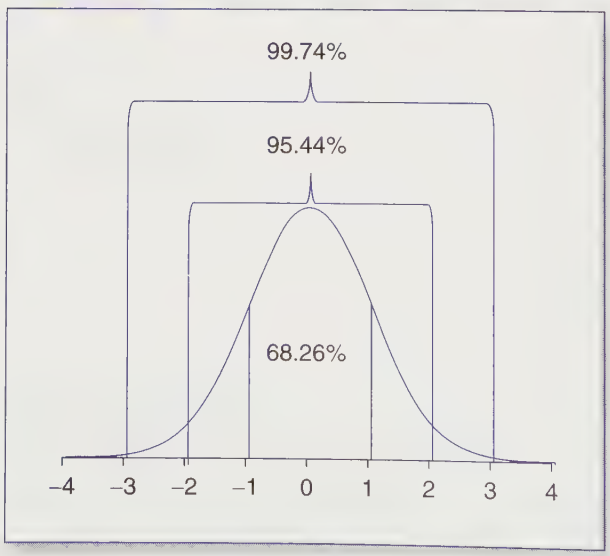
$$X = 30.40.$$

A raw score of 30.40 on this test represents performance that is 2 standard deviations above the population mean (i.e., very good performance!). This value could be used to identify very motivated students who score far above average.

z SCORES AND THE STANDARD NORMAL CURVE

We have learned that *z* scores are helpful because (1) they enable us to locate any score in a distribution of scores and (2) they provide a very systematic way to compare any score to any other score. For example, we know that positive *z* scores are above average and that positive *z* scores greater than +1 are further above the average than is typical of the positive *z* scores. And, we know that a *z* score of +0.38 is slightly better than a *z* score of +0.21. *z* scores always provide these two abilities to researchers. But, if the distribution of raw scores from which the *z* scores are derived is *normally shaped*, *z* scores enable a third, even more powerful ability to researchers. A normally shaped distribution of *z* scores enables researchers to make very precise probability statements about any score in a distribution. Before we explain how this third more powerful ability of *z* scores is possible, we need to discuss what we mean by a *normally shaped distribution*.

Figure 4.1 Percentage of Scores Between Standard Deviations in a Normal Distribution



When statisticians talk about a normally shaped distribution, or a normal curve, they mean a distribution of scores that has a *very* specific shape. Figure 4.1 displays a normal curve. First, normal distributions are symmetrical, meaning their left and right sides are identical so the mean, the median, and the mode are all the same value. Second, they are bell shaped, so they follow a 68–95–99 rule. In normal curves, 68.26% of all the scores in the distribution are between

1 standard deviation below the mean and 1 standard deviation above the mean (i.e., between −1 and +1 in Figure 4.1). In addition, 95.44% of all scores are between 2 standard deviations below the mean and 2 standard deviations above the mean (i.e., between −2 and +2 in Figure 4.1). Finally, 99.74% of the scores are between 3 standard deviations below and 3 above the mean (i.e., between −3 and 3 in Figure 4.1). From this rule, you can see that scores 2 or more standard deviations above the mean are *very rare* in a normal curve.

You might be thinking that this is very interesting, but how often do distributions actually conform to all of these specifications? More often than you might think. In actuality, it's called a "normal" curve because it occurs quite frequently. A great many variables when measured in a population will generate a normal curve. Furthermore, because of a theorem you will learn about in Chapter 5, the central limit theorem, normal curves are extremely common in research situations. You will read more on why in the next chapter. The key point is that normal curves are quite common and you should understand the 68–95–99 rule, that approximately 68% of scores are between the z scores +1 and −1, 95% are between +2 and −2, and 99% are between +3 and −3.

To return to our central point, when a distribution has a normal shape, we can use z scores to make very precise probability statements about any score in that distribution.

For example, you can determine what proportion of scores is higher or lower than a particular z score. So, if test scores were normally distributed, you could determine what proportion of people scored higher than (or lower than) you on that test. Similarly, if you randomly select one score from a distribution, you can determine the probability of selecting a z score of that size or larger from the distribution. Determining the probability of getting a specific z score at random, or by chance, will be extremely important when we learn about significance testing in Chapter 6.

Reading Question

7. z scores can be used to make precise probability statements about the location of a raw score relative to other scores only if the raw scores are normally distributed.

   a. True

   b. False

Reading Question

8. In a normal distribution of scores, 68.26% of the z scores fall between ____ and ____ standard deviations of the mean.

   a. −1 and +1

   b. −2 and +2

   c. −3 and +3

The first step to performing any of the above statistical procedures is to convert one or more raw scores into z scores. Any raw score can be converted into a z score simply by using the z for a single score formula (e.g., $z = (X - \mu)/\sigma$). If you convert *all of the raw scores* in a distribution into z scores, you will end up with a distribution that has a mean of 0 and a standard deviation of 1.

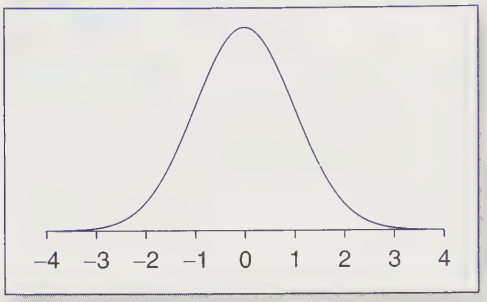
Reading Question

9. If you convert an entire distribution of raw scores into z scores, the distribution of z scores will have a mean equal to _____ and a standard deviation equal to _____.

   a. 1, 0

   b. 0, 1

Figure 4.2 A Frequency Histogram of *z* Scores for a Normally Shaped Distribution



As mentioned previously, although you can convert any set of scores into *z* scores, you can only determine the probabilities of scores if the original (or raw) scores are normally distributed. If the scores are normally distributed, the distribution of *z* scores will be a normal curve with a mean of 0 and a standard deviation of 1. Converting raw scores into *z* scores essentially "translates" the raw scores into a *standard* normal distribution that has a mean of 0 and a standard deviation of 1.

As you can see in Figure 4.2 and is implied by the 68–95–99 rule, in a standard normal curve, there are many *z* scores close to 0 (the mean *z* score) and fewer *z* scores that are far from 0. The "peak" of the curve is over 0 (the mean *z* score), and the height of the curve decreases as you move farther away from 0. Scores that are common (e.g., 0) have a lot of "area" under the curve. Conversely, scores that are rare (e.g., −3 and + 3) have very little area under the curve. It is possible to get *z* scores that are farther from 0 than −3 and +3, but as the curve implies, this is rare. In fact, the normal curve allows us to say *exactly* how rare.

Reading Question

10. In a normal distribution of *z* scores, the most common *z* score value is
    a. 1.
    b. −1.
    c. 3.
    d. −3.
    e. 0.

The following three problems exemplify how you can use *z* scores from a standard normal curve to determine the exact probability of any given score.

EXAMPLE 1: POSITIVE *z* SCORE

To make a precise statement about the location of a score in a distribution, begin by computing a *z* for a single score, as you did above. For example, suppose Harriet asks you to help her compare her ACT score to others' ACT scores. Your first step is computing the *z* score for her ACT score of 22. As you may recall, the population mean score on the ACT is $\mu = 21$, with a standard deviation of $\sigma = 4.70$.

Compute the *z* Score

This is exactly what you did at the beginning of this chapter.

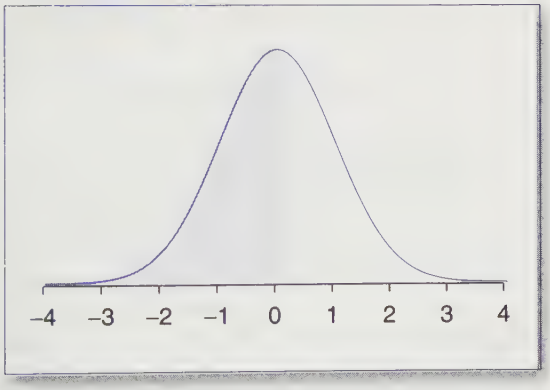
$$z = \frac{X - \mu}{\sigma} = \frac{22 - 21}{4.70} = 0.21.$$

From her z score of +0.21, Harriet knows that she scored slightly above the population mean. However, because the population of ACT scores is normally distributed, like many variables, you can describe her performance even more precisely. You can determine the proportion of ACT scores that were lower than her score of 22. In other words, she can find her percentile rank, *the percentage of scores that are equal to or lower than a score.*

Draw a Normal Distribution, and Shade the Area You Are Interested In

In this problem, you want to find the percentile rank for a z score of +0.21. You should (1) sketch a normal curve like that shown in Figure 4.2, (2) locate Harriet's z score on this curve and place a mark at that location, and (3) determine if the problem your trying to solve requires you to find the area to the left of Harriet's score or to its right. In this problem, we want to know the proportion of ACT scores that are equal to or less than 22, so we shade that area in the curve that is to the left of +0.21 on the z score number line. Figure 4.3 illustrates what your drawing should look like. After you have drawn a figure similar to the one shown in Figure 4.3, you determine the exact proportion of the curve's area that is shaded by using a unit normal table.

Figure 4.3 A z Distribution With the Target Area Below $z < 0.21$ Shaded



Use a Unit Normal Table (Located in Appendix A of This Book) to Find the Area of the Shaded Proportion of the Curve

A unit normal table allows you to look up any z score and determine the proportion of scores in the normal curve that are less than or more than that z score. So, you need to know which of these options the question is asking you to find. In this question, you want to know the percentage of scores that are equal to or less than her score of 22. To use a unit normal table, you need to refer to your sketch (e.g., Figure 4.3) and determine if the shaded area is more than or less than 50% of the curve. If the shaded area is less than 50% of the curve, you will need to use the "tail" column in the unit normal table. However, if it is more than 50%, you will use the "body" column in the table.

- If the shaded area is more than 50% of the distribution, use the body column in the table.
- If the shaded area is less than 50% of the distribution, use the tail column in the table.

Reading Question

11. If the area of the normal curve you want to find is more than half of the distribution, you should use the _____ column of the unit normal table to find its area.
    a. tail
    b. body

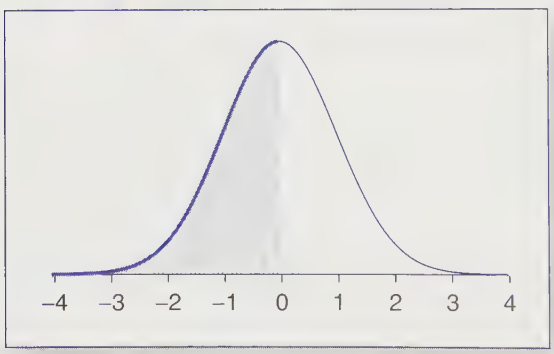
Finding Harriet's percentile rank is as easy as finding the z score of 0.21 in the z score column of Appendix A and reading the value next to that z score in the body column of the table. Again, we use the body column because the shaded area is more than half of the distribution. The body column indicates that the proportion of scores equal to or less than a z score of +0.21 is .5832. Therefore, 58.32% of the population of ACT scores were equal to or lower than Harriet's ACT score of 22. In other words, Harriet is at the 58.32nd percentile. This estimate of 58.32% is based on the assumption that the population of ACT scores is normally distributed. If the population distribution varies greatly from a normal shape, this estimate would be inaccurate.

EXAMPLE 2: NEGATIVE z SCORE

Antonio took the SAT. He wants to know the percentile rank of his combined math and verbal score on the SAT, which was 1005 (i.e., $X = 1{,}005$). The mean combined math–verbal score on the SAT is $\mu = 1008$, with a standard deviation of 114. Again, the first step is to compute the z score for a combined SAT score of 1005. This is done below:

$$z = \frac{X - \mu}{\sigma} = \frac{1005 - 1008}{114} = -0.026.$$

Figure 4.4 A z Distribution With the Target Area Below $z < -0.026$ Shaded



Draw a Normal Distribution, and Shade the Area You Are Interested In

The next step is to draw a normal distribution of combined SAT scores, locate the z score of −0.026, find the percentile rank of Antonio's score, and shade the area under the curve that is less than the z score of −0.026. This is done in Figure 4.4.

Use a Unit Normal Table to Find the Area That Is Shaded

For this example, you want to know the proportion of z scores that are *less than a negative z score* of −0.026. Note that the unit normal table does not include negative values. Therefore, you will always look up the absolute value of the z score you compute (i.e., in this case, you would look up 0.026) and then decide if you should use the body or the tail column based on your sketch. If the shaded area is more than half of the distribution, use the body column. If the shaded area is less than half of the distribution, use the tail column. In this case, less than half of the distribution is shaded, so we use the tail column for the z score of 0.026. The table in this book only goes two places past the decimal, so you should round 0.026 to 0.03. If you look up 0.03, you will find that the proportion of scores in the tail column is .4880. Therefore, approximately 48.80% of the students who took the SAT had combined math–verbal SAT scores that were equal to or lower than Antonio's score of 1005.

If you wanted to compare Antonio's SAT performance (i.e., 1,005) with Harriet's ACT performance (i.e., 22), the raw scores are not very helpful. However, Antonio's and Harriet's z scores of −0.026 and 0.21, respectively, make it clear that Antonio scored slightly below the SAT mean and that Harriet scored slightly above the ACT mean. If you wanted to be even more precise in your comparison, you could use the percentile ranks associated with each z score. Antonio's SAT performance was equal to or better than 48.80% of the SAT scores, and Harriet's ACT performance was equal to or better than 58.32% of the ACT scores. So, Harriet's ACT performance placed her approximately 10 percentile points higher in the ACT distribution than Antonio's SAT performance placed him in the SAT distribution.

Reading Question

12. When finding the area of the distribution that is above or below a negative z score, you
    a. need a different z table with negative values in it.
    b. simply look up the absolute value of the z score and then determine if you need to use the tail or body column of the table.

Reading Question

13. What percentage of z scores are below a z score of −.50? (Hint: Draw a normal curve, locate −.50 on that curve, and determine if you should use the body or tail column of the unit normal table.)
    a. .6915
    b. .3085

EXAMPLE 3: PROPORTION BETWEEN TWO z SCORES

Professor Jones wanted to know the proportion of students with ACT scores between 1 standard deviation above the mean (i.e., a z score of +1) and 1 standard deviation below the mean (i.e., a z score of −1). The first step in this problem is to draw the distribution of ACT scores and sketch in the area you are trying to find. This has been done below.

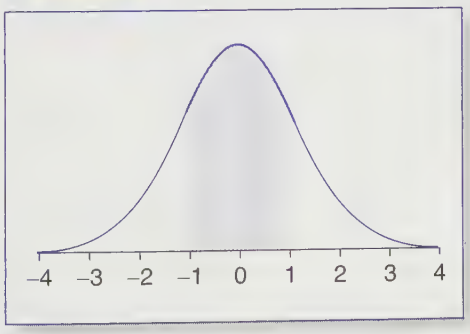


Figure 4.5 A z Distribution With the Target Area Between $z = -1$ and $z = +1$ Shaded.

Draw a Normal Distribution, and Shade the Area You Are Interested In

In this example, you are trying to find the proportion of students in the middle of the curve, with z scores between −1 and +1. The area between the z scores of −1 and +1 has been shaded in Figure 4.5.

Use a Unit Normal Table to Find the Area That Is Shaded

There are a number of correct ways to do this problem. One would be to determine the entire area greater than −1. Because this is greater than 50% of the distribution, we look in the body column for the z score of 1 and find .8413. However, we aren't interested in everything greater than 1. Specifically, we

want to exclude the area greater than the z score of +1, or the area in the tail for a z score of 1, which is .1587. Therefore, the area that we're interested in, the area between the z scores of −1 and +1, is .8413 − .1587 = .6826. Look at the figure above, and confirm that you understand why we had to subtract the area in the positive tail to get our answer.

The proportion of ACT scores between the z scores of −1 and +1 was .6826. Therefore, 68.26% of all ACT scores are between 1 standard deviation below the population mean and 1 standard deviation above the population mean. In other words, most people (approximately 68% of them) have ACT scores between 16.3 (i.e., 21 − 4.7) and 25.7 (i.e., 21 + 4.7).

Reading Question

14. If you want to determine the area of the normal curve that is between any two z scores, you will need to
    a. use a different normal unit table.
    b. sketch the area you need and then subtract one area from another.

OVERVIEW OF THE ACTIVITY

In Activity 4.1, you will compute z scores and determine the probability of obtaining those z scores using the unit normal table. When talking about the probabilities of particular z scores, you will also identify which outcomes are more or less likely than other outcomes.

Activity 4.1: z Scores and Probabilities

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Compute z scores for individual scores
- Determine the probability of a score (or z score) using a unit normal table
- Identify likely and unlikely raw scores by using z scores

PART I: z SCORES AND THE UNIT NORMAL TABLE

1. In this activity, you will compute z scores and use them to (1) locate scores in a distribution, (2) compare scores within the same distribution and across different distributions, and (3) determine the probability of specific z scores. The chapter reading made the point that you can always use z scores to accomplish (1) and (2), but only in a very specific circumstance can you accomplish (3) with z scores. What is that circumstance? You can only use z scores to determine the probability of a score if
   a. the z score is positive.
   b. the z score is negative.

c. the population distribution is normally shaped.

d. the population distribution is negatively skewed.

e. the population distribution is positively skewed.

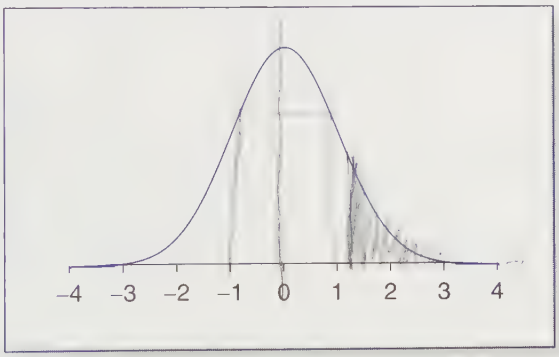
2. If the population distribution of scores is normally distributed, you can use the unit normal table to determine the probability of getting a z score. However, to use the table correctly, you have to know when a problem is asking for a "tail probability" or a "body probability." As indicated in the reading, you should sketch a normal curve, locate the z score, and determine if you need the probability of scores less than or greater than the z score. If the shaded area in your sketch is larger than half of the curve, use the _____ of the unit normal table.

a. body column

b. tail column

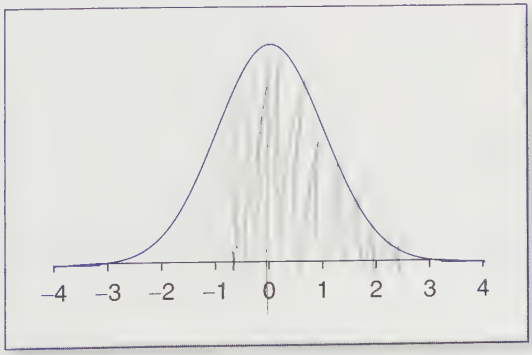
For Questions 3 to 6, determine if you need a body probability or a tail probability by creating a sketch of the problem on the provided normal curve. Then find the actual probability for each question by using Appendix A.

3. What proportion of z scores are equal to or greater than a z score of +1.2 on a normal curve?



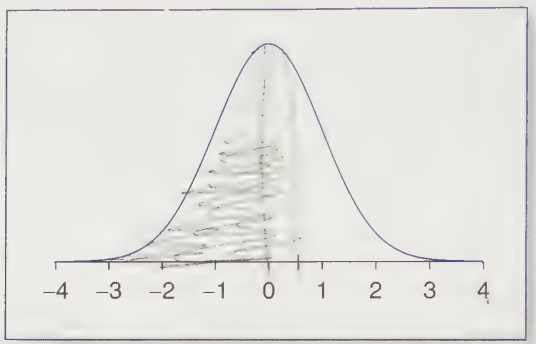
tail
.1151

4. What proportion of z scores are equal to or greater than a z score of −0.75 on a normal curve?

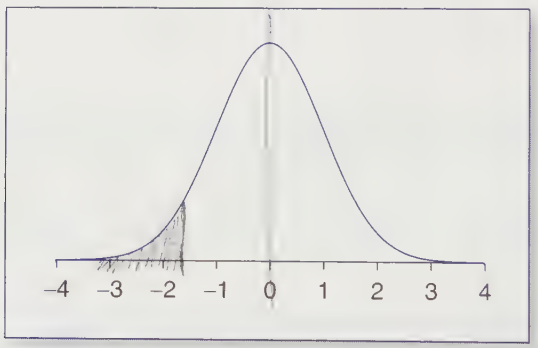


body
.7734

5. What proportion of scores are equal to or less than a z score of +0.66 on a normal curve?



6. What proportion of scores are equal to or less than a z score of −1.65 on a normal curve?



PART II: COMPUTING z SCORES AND FINDING PROBABILITIES

7. Unfortunately, Henry was in a motorcycle accident in which he suffered a serious head trauma. Dr. Gatz, a neurological psychologist, is giving Henry several cognitive tests to help determine if specific neurological pathways were damaged by the accident. For the first test, Henry had to listen to lists of words and then repeat them back. Dr. Gatz recorded how many words Henry could hear and then repeat back correctly. The adult population can repeat back an average of 7 words with a standard deviation of 1.30 words (i.e., $\mu = 7$; $\sigma = 1.30$). Henry was able to repeat back 6 words. Use a z score to determine the proportion of people in the adult population who scored better than Henry on this memory test. Assume that the population of memory scores on this test is normally distributed.

8. Next, Dr. Gatz tested Henry's ability to identify simple line drawings of objects. In the adult population, the parameters were $\mu = 10$; $\sigma = 0.89$. Henry named seven objects correctly. Use a *z* score to determine the proportion of people in the adult population who scored better than Henry on this object naming test. Assume that the population of object naming scores on this test is normally distributed.

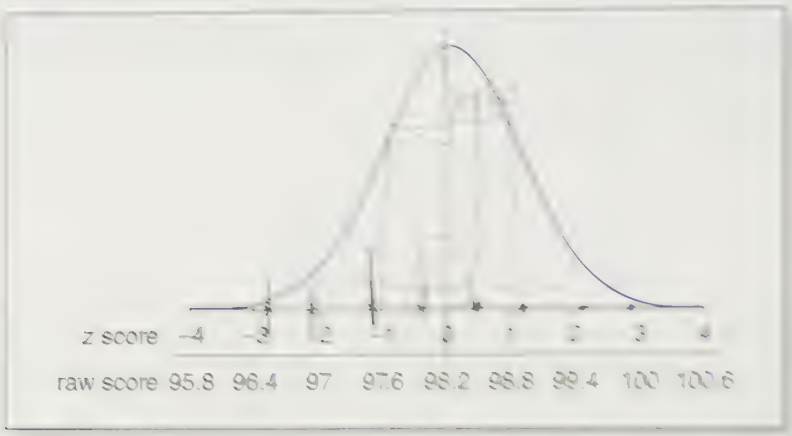
9. Finally, Henry performed the Stroop test. He was told to name the color of ink that a list of color names (e.g., blue, red, orange) were printed in. Each color name was printed in an ink color that was different from the color name (e.g., blue might be printed with black ink). The time it took Henry to name the ink color of all the words was recorded. *For this test, faster times indicate better performance.* It took him 15.70 seconds to complete the task. In the adult population, the parameters were $\mu = 16.20$ seconds and $\sigma = 1.30$ seconds. Use a *z* score to determine the proportion of people in the adult population who scored better than Henry on this test. Assume that the population of Stroop test times is normally distributed.

10. Now, Dr. Gatz needs your help to interpret the three *z* scores that you computed. Based on Henry's *z* scores of the memory test, the object naming test, and the Stroop test, do you think Dr. Gatz should focus his future investigations on Henry's memory systems, naming systems, or attentional systems (i.e., those assessed by the Stroop test)? Explain your reasoning to Dr. Gatz.

PART III: BODY TEMPERATURE

Almost all of us have had our body temperature measured at the doctor's office. Not only is temperature an indicator of physical health, but it is also associated with psychological health. For example, people with major depression tend to have higher body temperatures than people who are not depressed (Rausch et al., 2003). Conventional wisdom is that the average human body temperature is 98.6 °F. This value seems to have originated from Carl Wunderlich's 1868 publication, in which he says that he measured the body temperatures of approximately 25,000 patients. However, more recent research (Mackowiak, Wasserman, & Levine, 1992) has revealed that the average body temperature (taken orally) of healthy adults is $\mu = 98.2$, with a standard deviation of $\sigma = 0.60$. The following graph is a frequency distribution of body temperatures.

11. Locate the following individual body temperatures on the normal curve.



98.0 °F, 97.5 °F, 96.9 °F, 96.5 °F. Put an "x" on the number line of the graph for each temperature.

12. Rank the four body temperatures in order from most likely to least likely.

98, 97.5, 96.9, 96.5

13. How can you use the frequency distribution histogram of the population's body temperatures to determine which of the preceding temperatures are most rare? Select all that apply.

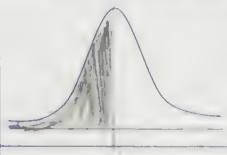
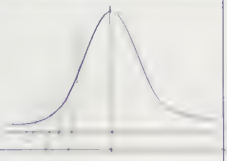
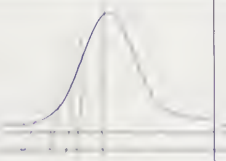
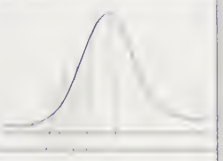
a. Scores that are the farthest from the mean are the most rare.

b. Negative *z* scores are more rare than positive *z* scores.

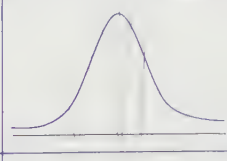
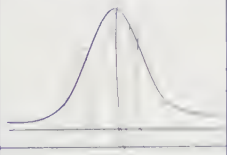
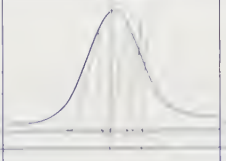
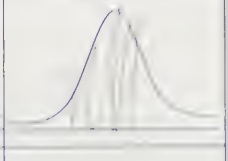
c. The height of the curve above a given score indicates how rare that score is: the lower the curve, the rarer the scores are at that location.

In the previous questions, you visually inspected a frequency distribution graph to determine whether certain body temperatures were unusual. While this procedure is informative, it is not very precise. You can use *z* scores to determine the exact probability of obtaining any particular range of scores in a distribution. As you know from the reading on *z* scores, you need to know the population mean (μ) and population standard deviation (σ) for all body temperatures to compute the *z* score for any given body temperature. We will use $\mu = 98.2$ and $\sigma = 0.6$.

14. Find *z* scores for the body temperatures that follow. Then, shade the area of the graph that includes all of the scores that are farther from the mean (i.e., more extreme) than the one you computed. Use Appendix A to find the probability of obtaining a *z* score that is equal to or more extreme (i.e., farther from zero, in this case lower) than the one you computed.

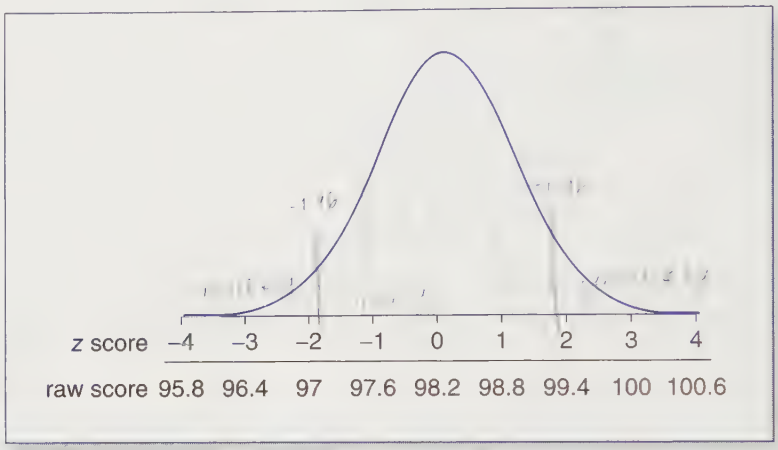
Test Score	98.0 °F	97.5 °F	96.9 °F	96.5 °F
$z = \frac{X - \mu}{\sigma} =$				
Shade the graph				
Probability of obtaining a *z* score equal to or lower than the *z* you computed				

15. Find *z* scores for the body temperatures listed below. Then, shade the area of the graph that includes all of the scores *higher* than the one you computed. Finally, use the *z* score table to find the probability of obtaining a *z* score that is equal to or more extreme (i.e., farther from zero, in this case higher) than the one you computed.

Test Score	98.4 °F	98.9 °F	99.5 °F	99.9 °F
$z = \frac{X - \mu}{\sigma} =$				
Shade the graph				
Probability of obtaining a *z* score equal to or lower than the *z* you computed				

Once the probability of a *z* score or a more extreme *z* score is found, you will still need to determine if that probability is small enough to be considered *unlikely*. Researchers frequently use a 2.5% cutoff, meaning that if the probability of obtaining a score that is as extreme or more extreme than a given score is less than 2.5%, it is considered rare or uncommon.

16. Find the *z* score cutoff that separates the top 2.5% from the rest of the distribution by looking up .025 in the tail column of Appendix A. After you find .025 in the tail column, find the *z* score on that row. The *z* score is 1.96. A positive *z* score of 1.96 is the cutoff that separates the top 2.5% from the rest of the distribution. A *z* score of −1.96 is the cutoff that separates the bottom 2.5% from the rest of the distribution. Draw a vertical line at each cutoff point.



17. Label each of the three sections that you created for the previous question as "likely scores" or "unlikely scores." Two of the sections will be labeled as "unlikely scores."

18. Find the body temperature (i.e., raw score) that corresponds to the positive *z* score cutoff of +1.96 and the body temperature that corresponds to the negative *z* score cutoff of −1.96 by using the *z* score formula ($z = \frac{X - \mu}{\sigma}$). You know *z*, μ, and σ; you need to solve for *X*.

The temperature cutoff for the *z* of +1.96 = 99.376

The temperature cutoff for the *z* of −1.96 = 97.024

19. Determining if a body temperature is unusual and possibly a sign of a problem can be difficult because there is quite a bit of individual difference variability in body temperatures. A physician could use *z* scores to determine if a body temperature is rare and may be a sign of a problem. Based on the *z* score cutoffs you defined above, which of the following body temperatures (from the previous problems) would be unlikely and should be investigated more closely? Use the *z* scores you computed in Questions 14 and 15 to answer this question. Circle all that are unlikely.

98.0 97.5 96.9 96.5 98.4 98.9 99.5 99.9

Chapter 4 Practice Test

1. Scores on an IQ test are normally distributed with a mean of 100 and standard deviation of 15. What is the z score for an IQ score of 122?
   a. −1.65
   b. 1.65
   c. −1.47
   d. 1.47

2. Scores on an IQ test are normally distributed with a mean of 100 and standard deviation of 15. What is the z score for an IQ score of 93?
   a. −.47
   b. .47
   c. .65
   d. −.65

3. Scores on an organic chemistry exam were normally distributed with a mean of 43 and a standard deviation of 7.6. What percentage of students had scores higher than 60?
   a. 98.75%
   b. 1.25%
   c. 9.875%
   d. .125%

4. Scores on an IQ test are normally distributed with a mean of 100 and a standard deviation of 15. To help identify people who need extra help in school, a school district needs to know the IQ for students who are in the bottom 10% of the IQ score distribution. What is the cutoff?
   a. 75.2
   b. 55.2
   c. 80.8
   d. 84.7

5. A math teacher gives students their exam grades back as z scores rather than as percentages or raw scores. A student gets his test back with a score of +2.2. How did this student do on the exam compared to the other students in the class?
   a. Great!
   b. About average
   c. Below average
   d. There is no way to know without knowing the mean. What sort of teacher gives test scores back as z scores?!

6. In a normal distribution, what percentage of z scores are between −1 and 1?
   a. 84.13%
   b. 15.87%
   c. 31.74%
   d. 68.26%

7. Scores on a measure of racism form a normal distribution with a mean of 50 and a standard deviation of 11. What would the mean and standard deviation be if all scores in the distribution were converted into z scores?
   a. Mean = 0, standard deviation = 1
   b. Mean = 1, standard deviation = 1
   c. Mean = 0, standard deviation = 0
   d. Mean = 1, standard deviation = 0

8. A student in a math class receives a z score of +1.5 on the first exam. How should she respond to this news?
   a. That's great! I must really know the material because I did better than over 93% of the students in the class.
   b. That's not so great. I only scored higher than about 6% of the students in the class.
   c. Well, I'm glad I did better than the class average by more than was typical, but I don't know how well I understood the material based on this score. Maybe we all did really poorly on the test and I just did better than average.
   d. Wow, we all really bombed this test!

9. Which of the following z scores is the most likely to occur in a normal distribution?
   a. 2.3
   b. 1.2
   c. −.3
   d. −1.9

10. If a distribution of scores is very skewed (i.e., not close to being normally distributed), which of the following cannot be done?
    a. Find the median
    b. Convert raw scores into z scores
    c. Use the unit normal table to identify the probability of scores
    d. All of the above can be done with a skewed distribution

REFERENCES

Mackowiak, P. A., Wasserman, S. S., & Levine, M. M. (1992). A critical appraisal of 98.6 degrees F, the upper limit of the normal body temperature, and other legacies of Carl Reinhold August Wunderlich. *Journal of the American Medical Association, 268,* 1578–1580.

Rausch, J. T., Johnson, M. E., Corley, K. M., Hobby, H. M., Shendarkar, N. N., Fei, Y. Y., . . . Leibach, F. H. (2003). Depressed patients have higher body temperature: 5-HT transporter long promoter region effects. *Neuropsychobiology, 47*(3), 120–127.

CHAPTER 5

The Distribution of Sample Means and *z* for a Sample Mean

LEARNING OBJECTIVES

After reading this chapter, you should be able to do the following:

- Explain how a distribution of sample means is created
- Explain how a random sample is obtained
- Determine the mean, the standard deviation, and the shape of a distribution of sample means
- Explain what the standard error of the mean measures
- Explain the central limit theorem and why it is important
- Explain the law of large numbers
- Compute a *z* for a sample mean
- Use the *z* for a sample mean and a unit normal table to determine how likely a given sample mean is to occur

SAMPLING AND SAMPLING ERROR

In most situations, researchers want to learn about entire populations, but in most situations, it is impossible to study entire populations because they are too large. Consequently, researchers are forced to study a sample and to infer that the sample results represent what they would find if they studied the entire population. You learned in Chapter 1 that *using sample statistics to infer population parameters* is called inferential statistics. While absolutely necessary, the problem with all inferential statistics is they introduce *sampling error*. Sampling error occurs when the sample's characteristics differ from the population's characteristics. The greater the discrepancy, the greater the sampling error. For example, if a population is 75% female with a mean age of 21, the sample used in an investigation of this population should also be 75% female with a mean age of 21. If a sample had dramatically different characteristics than the population, it would provide sample

statistics that are dramatically different from the population, creating a lot of sampling error. If you are trying to represent a specific population and get a poor sample, even if the rest of the study is executed perfectly, the sample results will not represent those of the population, and the entire purpose for doing the study is compromised. Therefore, obtaining a representative sample that minimizes sampling error is critically important. Researchers minimize sampling error with good sampling procedures.

Reading Question

1. Using inferential statistics creates which of the following problems for researchers?
   a. Statistical error
   b. Sampling error

When you take a research methods course, you will learn about the different sampling procedures researchers use to get representative samples that minimize sampling error. But, it is important that you understand the basics of sampling so that you recognize inferential statistics' limitations. Suppose Dr. Reminder wants to study how aging affects prospective memory (i.e., one's memory for doing future tasks like taking out the garbage before 8 a.m., taking a pill at 2 p.m., or going to an appointment at 4 p.m.). Dr. Reminder goes to a local senior citizen center and gets 25 people between the ages of 65 and 80 to agree to participate in his study next Tuesday at 3 p.m. However, only 10 of these 25 people remember to show up for the study. Clearly, concluding that the results from only those who remembered to show up represent the entire population, even those who forgot to show up, is problematic. Perhaps the people who remembered to show up have better prospective memory abilities than those who forgot to show up. In this case, Dr. Reminder's sampling procedure was poor, so he ended up with a nonrepresentative sample of aging people. Therefore, even if he did every other part of his study perfectly, his results from this sample will misrepresent what the population's results would likely be. The only way to fix this problem is to make the sample more representative by somehow getting more of the people with lower prospective memory abilities into the study's sample. The key point is that if poor sampling procedures created a poor sample, the entire study is compromised until the sampling procedure is fixed. You need to recognize that statistical results are only accurate if the sample is representative of the population.

Reading Question

2. One way researchers can address the problem of sampling error by using a good, nonbiased sampling procedure.
   a. True
   b. False

Experienced researchers take great care when obtaining samples because they recognize the critical importance of representative samples. However, even if you use a good sampling procedure, it is likely that you will still have sampling error in your study. Thus far, we have talked about sampling error only in general terms as the discrepancy between a sample statistic and a population parameter. In this chapter, you will learn to compute the expected amount of sampling error in a study and learn how to minimize sampling error.

Computing sampling error is relatively simple. You simply divide the standard deviation of the population by the square root of N Sampling error is also referred to as the standard error of the mean (*SEM*), and so the formula is

$$SEM = \frac{\sigma}{\sqrt{N}}.$$

Thus, there are only two things that contribute to sampling error: the standard deviation of the population and the size of the sample. As σ increases, the *SEM* also increases. As the sample size increase, the *SEM* decreases. In general, researchers have far less control over the population standard deviation than the sample size, and so efforts to reduce sampling error tend to focus on increasing the size of the sample. For example, suppose that you know that scores on a reading comprehension test have a population mean of 500 with a standard deviation of 100. You want to do a study of reading comprehension and are considering using a sample of 10 people and want to know how much sampling error you could expect. To figure this out, you would just divide σ by $\sqrt{N}$:

$$SEM = \frac{\sigma}{\sqrt{N}} = \frac{100}{\sqrt{10}} = 31.62.$$

This simple computation tells you that if you did your study with a sample of 10 people, the typical distance you would expect between your sample mean and the true population parameter would be 31.62.

What if you doubled your sample size to 20? In this case, your sampling error would decrease substantially from 31.62 to 22.36.

$$SEM = \frac{\sigma}{\sqrt{N}} = \frac{100}{\sqrt{20}} = 22.36.$$

If you estimate your expected sampling error with increasingly large samples, you would find that as your sample size increases, sampling error decreases (see Table 5.1).

Although reduced sampling error is important, you must also weigh the costs and diminishing returns of obtaining increasingly large samples when deciding on your study's sample size. In this example, increasing the sample size by 30 people from 10 to 40 cuts the sampling error in half. However, to cut sampling error in half again, you would have to add another 120 people to the sample. If you inspect Table 5.1 carefully, you will notice that increasing the sample size by a factor of 4 decreases the expected sampling error by about half. Choosing an appropriate sample size for a study is a relatively complex topic influenced by a number of factors that you will learn about in a research methods course. For right now, understand that as sample sizes increase, sampling error decreases.

Table 5.1 Relationship Between Sample Size and Sampling Error With σ = 100

Sample Size	*Sampling Error*
10	31.62
20	22.36
40	15.81
80	11.18
160	7.91
320	5.59
640	3.95
1,280	2.80
2,560	1.98

DISTRIBUTION OF SAMPLE MEANS

Computing the amount of sampling error expected in a study isn't difficult. You simply divide σ by the square root of the sample size (*N*). However, to understand why this simple sampling error formula works, you must understand distributions of sample means. A **distribution of sample means** *is the population of all possible random sample means for a study conducted with a given sample size.* In other

words, if a study was planned to have a sample of 25 people drawn from a large population of 100,000, the distribution of sample means for this study would include *the mean from every possible combination of 25 people from the population*. Most of the statistics you will learn about in this book depend on your understanding the distribution of sample means. There is little doubt that you can memorize the definition, but fully understanding what the set of all possible random sample means for samples of a given size means requires a bit more work.

To illustrate what a distribution of sample means is, we will work with a very small population of four scores: 10, 12, 14, 16. You want to create a distribution of all possible sample means you could obtain if you did a study with a sample size of $N = 2$. To do that, you must first create a list of all possible combinations of two scores from the population. When obtaining random samples, it is important to sample with replacement. Thus, each score has an equal probability of being selected for the first score in a sample as well as the second score in a sample. So, the score 10 must be paired with every score, including itself. The score 12 must be paired with every score, 14 with every score, and 16 with every score. If you generate this complete list of possible samples, you will have all 16 of the possible random samples when $N = 2$ from the population of four scores. All 16 possible samples, labeled A through P, are provided in Table 5.2. The second and third columns contain the two scores in each of the 16 samples, and the fourth column is the mean of the two

Table 5.2 Distribution of Sample Means for Population of Four Scores and a Sample Size of 2

Sample	*First Score*	*Second Score*	*Sample Mean*
A	10	10	10
B	10	12	11
C	10	14	12
D	10	16	13
E	12	10	11
F	12	12	12
G	12	14	13
H	12	16	14
I	14	10	12
J	14	12	13
K	14	14	14
L	14	16	15
M	16	10	13
N	16	12	14
O	16	14	15
P	16	16	16

Distribution of sample means:
All possible random samples of a size of 2 from a population of four scores (10, 12, 14, 16)

scores in each sample. This final column is the distribution of sample means; it is the list of all sample means that are possible when the sample size is 2.

The mean for each random sample of two scores is in the last column. This is the distribution of sample means. It is the set of all possible random samples of a size of 2 taken from this population of four scores. The distribution of sample means is displayed graphically in Figure 5.1.

Reading Question

3. Which of the following would create a distribution of sample means?
   a. Taking all possible combinations of people from a population and computing their mean
   b. Taking 100 samples from a population and computing their mean
   c. Taking all possible combinations of four people (i.e., all samples of 4) from a population and computing each sample's mean

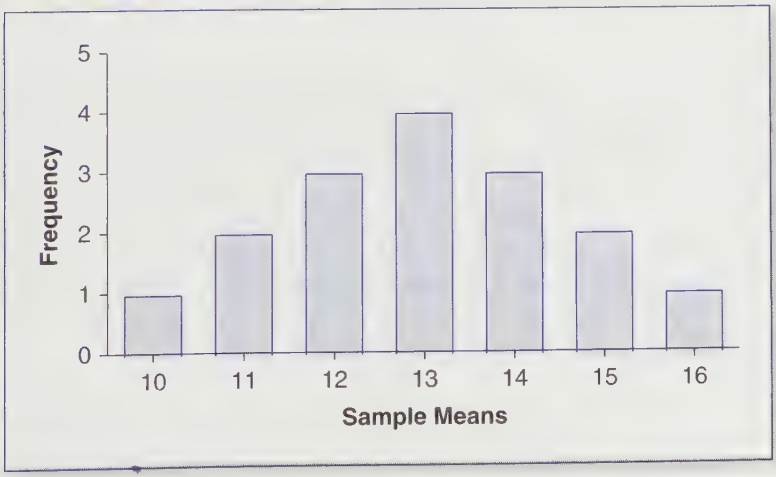
Now you can start to appreciate the special relationships between the original population of scores and the distribution of sample means. You started with a population of four scores: 10, 12, 14, and 16. The mean of this population was

$$\mu = \frac{\Sigma X}{N} = \frac{52}{4} = 13.$$

The mean of the distribution of sample means, the 16 sample means in the last column of Table 5.2 (i.e., 10, 11, 12, 13, 11, etc.), is also 13, as illustrated below:

$$\frac{\Sigma X}{N} = \frac{208}{16} = 13.$$

Figure 5.1 Distribution of Sample Means for Samples of 2 From a Population of Four Scores



Thus, the mean of the distribution of sample means is equal to the mean of the population. This will always be true. The fact that the mean of the distribution of sample means is always equal to the mean of the original population shows that when researchers randomly select a sample to use in their study, selecting a sample that has the same mean as the population, or a mean close to it, is fairly common. The frequency bar graph of the distribution of sample means in Figure 5.1 makes it easy to see why. The sample means that are close to the population mean of 13 are common, and the sample means far from the population mean are relatively rare.

In other words, sample means with low amounts of sampling error are common, and those with lots of sampling error are rare.

Another noteworthy aspect of the relationship between the original population of scores and the distribution of sample means is the systematic relationship between the standard deviation of the population and the standard deviation of the distribution of sample means. The standard deviation of the original population of scores (i.e., 10, 12, 14, 16) was

$$SS = \sum X^2 - \frac{(\sum X)^2}{N} = 696 - \frac{52^2}{4} = 20.$$

$$\sigma = \sqrt{\frac{SS}{N}} = \sqrt{\frac{20}{4}} = 2.24.$$

If we compute the standard deviation of the distribution of sample means, it will reveal the typical distance between the possible sample means and the population mean. This is the measure of sampling error. Below are the computations for the standard deviation of the distribution of sample means, beginning with the *SS*. Note that the *N* in the *SS* equation refers to the number of sample means (i.e., 16).

$$SS = \sum X^2 - \frac{(\sum X)^2}{N} = 2{,}744 - \frac{208^2}{16} = 40.$$

Compute the standard deviation of the distribution of sample mean by dividing the *SS* by *N*.

$$\sqrt{\frac{SS}{N}} = \sqrt{\frac{40}{16}} = 1.58.$$

This 1.58 is the typical amount of sampling error you should expect in your study. Most of the sample means that are possible will be within 1.58 of the population mean. In this example, we computed the standard error of the mean (*SEM*) by first creating a distribution of sample means and then finding the standard deviation of all the possible sample means. However, in virtually all research situations in the social sciences, you must work with much larger populations and sample sizes; consequently, it is impossible to actually create a list of all possible sample means (i.e., the distribution of sample means) as we did in this example. Fortunately, you don't have to. Instead of creating a list of all possible samples and then finding their standard deviation, you can use the simple formula we introduced at the beginning of the chapter to compute the typical sampling error of your study, called the *SEM* or standard error of the mean:

$$SEM_p = \frac{\sigma}{\sqrt{N}}.$$

The standard deviation for the population (i.e., 10, 12, 14, 16) was 2.24. The *N* refers to the size of the sample, not the number of samples in the distribution of sample means. Thus, the *SEM* is

$$SEM_p = \frac{\sigma}{\sqrt{N}} = \frac{2.24}{\sqrt{2}} = 1.58.$$

Amazingly, this value is identical to the standard deviation of the 16 sample means in the distribution of sample means. Of course, the advantage to this formula is that all you need to know is the standard deviation of the population and the size of the sample. This formula is an extremely simple way to estimate the sampling error of a study.

You now know that the mean of the distribution of sample means is always equal to the populations mean, μ, and the standard deviation is always equal to the *SEM*, $\frac{\sigma}{\sqrt{N}}$. The last thing you need to know about the distribution of sample means is its shape. If you create a graph of the distribution of sample means, as was done in Figure 5.1, you will see that it is approximately normally distributed (i.e., bell shaped). You can also look at the graph and see that the sample means clump up around the population mean of 13. All distributions of sample means will have a normal shape if the original population of scores has a normal shape. In this case, our population of four scores did not have a normal shape, but the distribution of sample means did approximate a normal shape. In general, as the sample size increases, the distribution of sample means will be more normal. If the original population is just mildly skewed, the distribution of sample means will be normally distributed with a fairly small sample size (e.g., 10), but when the original population of scores is highly skewed or has extreme outliers, you need a larger sample size (e.g., 40) to obtain a normal distribution of sample means. Because we often don't know what the original population of scores looks like, statisticians use 30 as a rough estimate of the sample size necessary to obtain a normal distribution of sample means. This normality characteristic of the distribution of sample means is important because normal curves enable us to make probability statements about any value in that normal curve, much like we did in the previous chapter.

You now know three things about the distribution of sample means, the set of all possible random samples of a given size from a population. First, this distribution has a mean equal to the population mean (μ). Second, the standard deviation is equal to the standard error of the mean $\left(SEM_p = \frac{\sigma}{\sqrt{N}}\right)$, a measure of sampling error. Third, the distribution is normal in shape as long as the sample size is at least 30 or the original population has a normal shape.

Collectively, these three facts are known as the central limit theorem (CLT), and they tell you the shape (normal), center (μ), and spread $\left(SEM_p = \frac{\sigma}{\sqrt{N}}\right)$ of the distribution of sample means for any study. Table 5.2 displays these the characteristics of the CLT. The CLT is powerful because (1) it suggests that samples tend to have means similar to the population from which they were drawn and (2) it enables us to compute the typical amount of sampling error any study is likely to generate.

Reading Question

4. The mean of the distribution of sample means is _____, and the standard deviation of the distribution of sample means is _____.

   a. the population mean; the population standard deviation

   b. the standard error of the mean; the population standard deviation

   c. the population mean; the standard error of the mean or the typical amount of sampling error a study is likely to have

Table 5.3 Three Components of the CLT and What They Do for Researchers

Characteristics of Any Distribution of Sample Means	*What Characteristic Does for Researchers*
Spread (standard deviation) will always be equal to $SEM_p = \frac{\sigma}{\sqrt{N}}$	Knowing the *SEM* tells researchers how much sampling error to expect
Center (mean) will always be the population mean	Knowing the mean tells researchers what sample mean to expect
Shape will be normal if the original population's shape is normal OR **Shape** will *approach* normal if the sample size of each sample is sufficiently large (e.g., $N \geq 30$)	Knowing the shape is normal enables using a normal unit table and determining the probability of any of a study's possible sample means

Reading Question

5. The shape of a distribution of sample means will
   a. have the same shape as the population.
   b. approach a normal shape as the sample size increases.
   c. approach a normal shape as the sample size decreases.

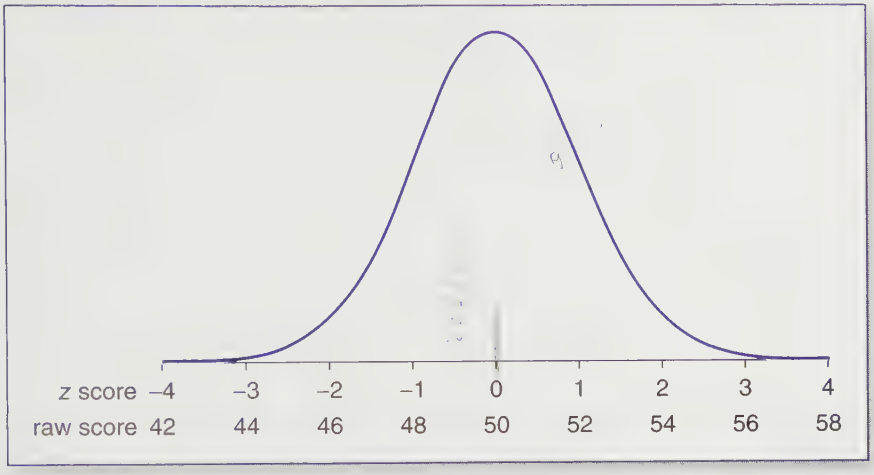
Reading Question

6. The central limit theorem only describes a distribution of sample means for populations of scores that are normally distributed.
   a. True
   b. False

Let's apply what you know about the CLT to Dr. Reminder's future study of prospective memory. Dr. Reminder wants to know how much sampling error his planned study would likely create when the population's mean is $\mu = 50$, its standard deviation is $\sigma = 10$, and his planned sample size is $N = 25$. By using the CLT, Dr. Reminder can create the distribution of sample means without computing every possible sample mean directly. Because of the CLT, he knows that the most common sample mean will be equal to the population mean, 50. He also knows that the typical amount of sampling error for his planned study will be equal to the standard error of the mean, $SEM_p = \frac{\sigma}{\sqrt{N}} = \frac{10}{\sqrt{25}} = 2$. Finally, Dr. Reminder knows that the distribution of sample means will have a normal shape because he knows that the population of scores on the prospective memory test, the original population, has a normal shape. With all this information, Dr. Reminder can create the distribution of sample means shown in Figure 5.2.

The fact that the mean of this distribution of sample means is 50, the population mean, tells Dr. Reminder that most of the possible sample means are close to the population mean (i.e., most have little sampling error) and that the study's most likely sample mean is the population mean.

Figure 5.2 A Distribution of Sample Means From a Population With a Mean of 50, a Standard Deviation of 10, and a Sample Size of 25



These facts are great news for the inferential statistics approach because they suggest that using samples to represent populations will work well most of the time. Notice how the most common sample mean (the peak of the frequency curve) is at 50, the population mean, and that sample means become less common the more distant they are from the population mean of 50. In other words, samples with lots of sampling error are rare.

Another important characteristic of the study's distribution of sample means is its standard deviation. The standard deviation of a distribution of sample means is the typical distance all the sample means deviate from the population mean. In other words, it is the typical amount of sampling error this study is expected to create. In fact, this characteristic is so important that it has a special name; the standard deviation of the distribution of sample means is called the **standard error of the mean** because *it measures the standard or typical amount of sampling error expected in a given study*. In this distribution of sample means, the standard error of the mean is 2. Some sample means will be closer than 2 away from 50 (i.e., have less sampling error), and some will be farther than 2 away (i.e., have more sampling error), but the typical sampling error of all means is 2, the standard error of the mean.

Reading Question

7. The standard error of the mean (SEM_p) is the
   a. typical discrepancy between all possible sample means of a given size and the population mean.
   b. discrepancy between a single score and a sample statistic.

Reading Question

8. The typical amount of sampling error is the
   a. standard deviation of the distribution of sample means.
   b. standard error of the mean.
   c. Both of the above

It should be clear that researchers want the standard error of the mean to be small because they want there to be as little sampling error as possible. Based on the sampling error formula $\left(SEM_p = \frac{\sigma}{\sqrt{N}}\right)$, you can deduce that there are two ways to reduce sampling error. One option is to reduce the population standard deviation (σ), but this option is rarely possible. You can't usually make scores less variable. A much more feasible option is to increase the size of the sample. As N increases, the overall standard error of the mean will decrease. Larger samples will tend to have means that deviate less from the population mean than smaller samples. This is a specific example of the **law of large numbers:** *As* N *increases, the sample statistic (e.g., the sample mean) is a better estimate of the population parameter (e.g., the population mean).*

Now Dr. Reminder can take advantage of the fact that this study's distribution of sample means is normally distributed to help him assess the expected sampling error of his planned study. For this study, the standard deviation of the distribution of sample means is 2. Because the distribution is normally shaped, it follows the 68–95–99 rule that you learned in Chapter 4. Therefore, 68% of all possible sample means are between −1 and +1 standard error of the mean of the population mean. Specifically, because the population mean is 50 and the standard error of the mean is 2, 68% of all possible sample means based on 25 people are between 48 and 52. Dr. Reminder can now use this information to determine if having a 68% chance of getting a sample within ± 2 points of the population mean is good enough or if he should increase his sample size to reduce this expected sampling error. Suppose Dr. Reminder decided that he wanted less expected sampling error than 2. He could increase his sample size from $N = 25$ to $N = 100$. This larger sample size would yield an expected sampling error of $SEM_p = \frac{\sigma}{\sqrt{N}} = \frac{10}{\sqrt{100}} = 1$. By increasing sample size from 25 to 100, Reminder cut the expected sampling error in half from 2 to 1. This is a good rule of thumb for you to know. Quadrupling sample size (i.e., $4 \times N$) cuts expected sampling error in half.

You learned in the previous chapter that researchers can use z scores and the unit normal table to determine the probability of getting any given score in a population of scores. Now you will learn that researchers can use a similar procedure, the z for a sample mean, to determine the probability of getting any given *sample mean* in a distribution of sample means.

Reading Question

9. Which of the following will *reduce* the standard error of the mean?
   a. Increasing N
   b. Decreasing N
   c. Both increasing and decreasing N can reduce the standard error of the mean.

Reading Question

10. According to the law of large numbers, as sample size increases, the sample mean tends to get closer to the population mean.
    a. True
    b. False

z FOR A SAMPLE MEAN

The CLT allows you to compute the standard error of the mean, which is the expected sampling error of a planned study. It also tells you that the distribution of sample means is frequently normally

shaped. Practically, this means that you can compute z scores for sample means, locate those means in a normal distribution, and use a unit normal table to determine the probabilities associated with any given sample mean. This statistical procedure enables you to do more than just compute expected sampling error. It enables you to make important decisions in the real world *based on statistical evidence*.

The formula for a z for a sample mean is very similar to the formula for a z for a single score. Both formulas are below.

$$z \text{ for an individual score formula: } z = \frac{X - \mu}{\sigma}.$$

$$z \text{ for a sample mean formula: } z = \frac{M - \mu}{SEM_p}, \text{ where } SEM_p = \frac{\sigma}{\sqrt{N}}.$$

For both formulas, the numerator is simply the observed difference between the score and a value. The numerator of the z for an individual score is the difference between an individual score and the population mean ($X - \mu$), while the numerator for the z for a sample mean is the difference between a sample mean and the population mean ($M - \mu$). For both formulas, the denominator is the typical amount of expected sampling error. For the individual score formula, σ is the typical distance of all possible individual scores from the population mean. For the sample mean formula, the standard error of the mean (SEM_p) is the typical distance of all possible *sample means* from the population mean. Thus, for both formulas, the obtained z score is a ratio of the observed difference over the difference expected due to sampling error. For both, a z score close to zero means that the observed deviation (i.e., the numerator) is small compared with the deviation expected by sampling error (i.e., the denominator). In other words, if the z score is "close" to zero, the score (i.e., X) or the sample mean (i.e., M) is "close" to the population mean. However, a z score of 3 means that the observed deviation (i.e., the numerator) is three times as large as the deviation expected by sampling error (i.e., the denominator). In other words, if the z score is far from zero, the score (i.e., X) or the sample mean (i.e., M) is "unexpectedly far" from the population mean. Large z scores often indicate that something other than sampling error variability caused the score (i.e., X) or the sample mean (i.e., M) to be very far from the population mean.

Reading Question

11. When computing a z for a sample mean, if the z score is close to zero the sample mean is
    a. not very different from the population mean and the difference is probably due to sampling error.
    b. very different from the population mean and the difference is probably created by something other than sampling error.

Reading Question

12. When computing a z for a sample mean, if the z score is far from zero, the sample mean is
    a. not very different from the population mean and the difference is probably due to sampling error.
    b. very different from the population mean and the difference is probably created by something other than sampling error.

EXAMPLE: COMPUTING AND INTERPRETING THE *z* FOR A SAMPLE MEAN

The *z* for a sample mean enables you to make important real-world decisions based on statistical evidence. For example, suppose you are a federal official investigating 16 loan officers at a large banking conglomerate you suspect of illegal mortgage practices. All loan officers at this bank were paid partly on commission, and so illegal activities could raise their income. By looking at the company records, you find that the mean income of all loan officers in the company is $\mu = 50{,}000$ with a standard deviation of $\sigma = 12{,}000$. Furthermore, the mean income for the 16 loan officers suspected of cheating is $55,000. You want to know the probability of these loan officers having a sample mean of $55,000 or greater due to chance (i.e., sampling error). Basically, you are asking, "What is the probability that the suspected employees' mean income was $55,000 (i.e., $5,000 higher than the company mean income) merely because of a coincidence (i.e., due to sampling error)?"

Step 1: Compute the Observed Deviation

Find the deviation between the sample mean (M) and the population mean (μ).

$$(M - \mu) = (55{,}000 - 50{,}000) = 5{,}000.$$

Step 2: Compute the Deviation Expected by Sampling Error

Next, compute the standard error of the mean.

$$SEM_p = \frac{\sigma}{\sqrt{N}} = \frac{12{,}000}{\sqrt{16}} = 3{,}000.$$

In this case, the standard error of the mean is 3,000. This means that when $N = 16$, the typical distance that all possible sample means deviate from the population mean is 3,000. When the sample size is 16, $3,000 is the amount of deviation we expect between any sample mean and the population mean due to sampling error.

Step 3: Compute the Ratio Between Observed and Expected Deviation (*z* for a Sample Mean)

To compute the *z* for a sample mean, divide the observed mean difference by the expected amount of sampling error (i.e., the standard error of the mean):

$$z = \frac{M - \mu}{SEM_p} = \frac{55{,}000 - 50{,}000}{3{,}000} = 1.67.$$

The *z* of 1.67 tells us that the observed deviation was 1.67 times greater than the deviation expected by sampling error.

Step 4: Locate the *z* Score in the Distribution

Once you have computed the *z* score (i.e., 1.67), you locate it on a *z* curve and determine if you need to look up a tail or body probability from Appendix A. You need to consider the question being asked. In

this situation, you want to know the probability of obtaining a sample of 16 people with a mean of $55,000 *or higher* due to sampling error. Therefore, you want to know the tail probability for a *z* score of 1.67 or higher; this is illustrated in Figure 5.3.

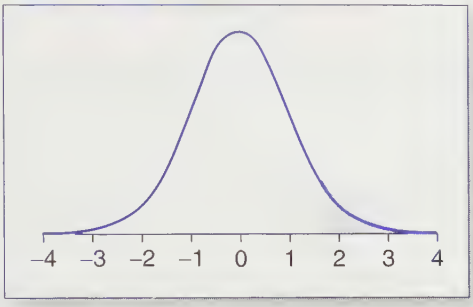


Figure 5.3 A *z* Distribution With the Target Area Below $z > 1.67$ Shaded

Step 5: Look Up the *z* Score

Use the unit normal table to determine the probability of obtaining a *z* score of 1.67 or higher due to sampling error. The area of the distribution above 1.67 is the smaller part of the distribution (i.e., less than half of the distribution), and so we look at the tail column for the *z* of 1.67 and find it is .0475.

Step 6: Interpret the *z* Score

When the sample size is 16, the probability of getting a sample mean income of $55,000 or greater is .0475. Another way of saying this is that 4.75% of all possible sample means based on 16 people have a mean equal to or greater than $55,000. It is *possible* that the salaries of the 16 suspects were higher than the company mean income merely due to chance or coincidence, *but it is not likely*. Specifically, we would only expect to obtain a random sample mean of $55,000 or higher 4.75% of the time. Therefore, the fact that the mean income of the 16 suspects was higher than the company's mean income was unlikely to be due to chance or sampling error. Perhaps the employees' salaries were higher due to illegal activities. Using your newfound statistical knowledge, you helped secure a warrant for the search of these 16 suspects' financial records. Nice work!

Reading Question

13. This example illustrates that the *z* for a sample mean procedure
    a. can be used to identify if the discrepancy between a sample mean and a population mean is likely or unlikely to be due to sampling error.
    b. is only useful in a controlled laboratory setting.
    c. Both a and b are correct.

EXACT PROBABILITIES VERSUS PROBABILITY ESTIMATES

The previous example illustrates that the *z* for a sample mean statistical procedure can be useful outside of controlled laboratory settings. There are many situations in which an investigator of some kind might want to know the probability that something is happening due to chance or sampling error. The central limit theorem enables us to compute the sampling error we expect. We can then compute the obtained *z* score and determine the probability of obtaining a score that extreme or more extreme due to chance. The central limit theorem is based on random sampling. However, researchers are rarely able to obtain truly random samples. Therefore, researchers are most often working with nonrandom samples. The result of working with nonrandom samples is that any probability statements are *estimates* rather than exact probabilities. As we matriculate through this course, try to remember that most (if not all) probability statements derived from research are *estimates*, and as a result, the actual probability might be a bit higher or a bit lower. Although researchers are usually creating probability estimates, the estimates are usually very good estimates.

Reading Question

14. The probabilities produced by most research studies are considered estimates (i.e., not exact probabilities) because
    a. most research studies do not use truly random samples.
    b. most research studies use samples that are too small.

OVERVIEW OF THE ACTIVITIES

In the first half of Activity 5.1, you will work with a very small population of four scores and generate a distribution of sample means by hand. You will also compute the mean and standard deviation of the distribution of sample means by hand and then compare these values to what is predicted by the central limit theorem (CLT). In the second half of the activity, you will work with a Java Applet that will allow you to look at a variety of different distributions of sample means and test the predictions of the CLT. In Activity 5.2, you will use the CLT and z scores to determine how likely you are to obtain sample means in a given range. Finally, you will distinguish between situations that require using a z for a single score versus a z for a sample mean.

Activity 5.1: Introduction to Distributions of Sample Means

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Describe a distribution of sample means
- Explain how a distribution of raw scores is different from a distribution of sample means that is created from those raw scores
- Find the mean and the standard deviation of a distribution of sample means
- Explain what the standard error of the mean measures
- Compute the sampling error
- Describe how the standard error of the mean can be decreased
- Explain why you would want the standard error of the mean to be minimized
- State the central limit theorem and why it is important

DISTRIBUTION OF SAMPLE MEANS AND SAMPLING ERROR

1. In Chapter 1, we introduced the idea that sample statistics are rarely exactly equal to the population parameters they are estimating. What is this difference between sample statistics and population parameters called?
   a. Standard deviation of the scores
   b. Sampling error

2. Suppose that the average height of women in the United States is 64.5 inches. If you obtained a random sample of 25 women from this population, what is your best guess as to what the mean height will be for that sample?

64.5in

3. In general, the _____ the sample, the closer the sample statistic should be to the population parameter.
   a. larger
   b. smaller

4. According to the law of large numbers, which of the following options is more likely? (Note: The mean height of women in the United States is 64.5 inches.)
   a. A random sample of two females from the U.S. population will have an average height of 70 inches.
   b. A random sample of 200 females from the U.S. population will have an average height of 70 inches.

5. Although it is important to understand the idea that larger samples result in less sampling error (i.e., the law of large numbers), you also need to be able to quantify the amount of sampling error for any given study. To do this, you will need to understand the distribution of sample means. Which of the following is the best description of a distribution of sample means?
   a. All possible sample means taken from a population
   b. The set of means for all possible random samples of a given size (N) taken from a population

6. According to the central limit theorem, the mean of the distribution of sample means will be equal to what value?
   a. 0
   b. μ
   c. σ

7. According to the central limit theorem, the standard deviation of the distribution of sample means will be determined by what formula?
   a. $\frac{\sigma}{\sqrt{N}}$ =SEM
   b. $\frac{SS}{N-1}$ - variance
   c. $\frac{\Sigma X}{N}$ = mean

8. According to the central limit theorem, the distribution of sample means
   a. is very often normally distributed.
   b. will always have the same shape as the original population of scores.

ILLUSTRATING THE DISTRIBUTION OF SAMPLE MEANS AND SAMPLING ERROR WITH A SMALL POPULATION AND SAMPLE

Being able to recite the definition of the distribution of sample means is a good start, but you need a far deeper knowledge of this topic if you hope to understand the more advanced topics in this course. In this activity, you are going to apply the CLT to predict the mean and standard error of the mean of a distribution of sample means for a very small population. Then you will actually compute every possible sample mean that the very small population could produce to confirm that the CLT's predictions were accurate. Researchers are usually interested in much larger populations, but it is much easier to illustrate what a distribution of sample means is by working with a very small population. The purpose of this first example is to illustrate that the CLT works; it predicts the mean and the standard error of the mean *perfectly*.

Suppose there is a very small population of four billionaires in a particular city. Further suppose that the following data represent the number of years of college/graduate school each billionaire completed:

2, 4, 6, 8 (Note: These data are completely made up.)

9. What is the mean for this population? $\mu =$ 5.

10. What is the standard deviation for this population? $\sigma =$ 2.36.

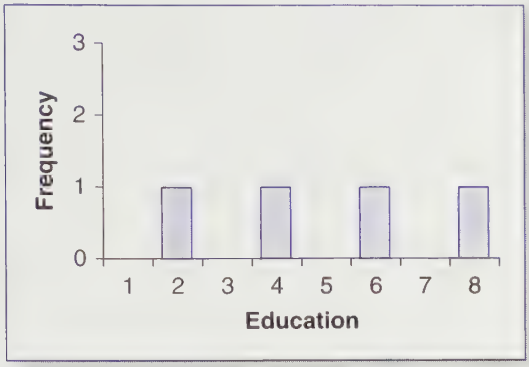
11. Now that you know the μ and the σ, the only other thing you need before you can apply the CLT is to decide on a sample size (N). Again, for illustration purposes, we will use an unusually small sample size of $N = 2$. According to the central limit theorem, the mean of any distribution of sample means will always be the pop. mean. (two words)

12. In this case, the predicted **mean of the distribution of sample means** will equal 5. (The answer is a number.)

13. Furthermore, according to the central limit theorem, the standard deviation of the distribution of sample means, or the **standard error of the mean,** will be $SEM_p = \frac{\sigma}{\sqrt{N}}$. So, in this case, the predicted **standard error of the mean** will equal 1.581. Now that you've predicted the center and spread of the distribution of sample means for the billionaire example when the $N = 2$, its time to determine if these predictions are *perfectly* accurate. (Hint: They will be!)

First, let's look at the shape of the original population, or parent population.

14. There is just one 2, one 4, one 6, and one 8, so the shape of this population's distribution is
(a.) not normal.
b. normal.



To create the distribution of sample means, you need to obtain the means for all possible random samples of sample size $N = 2$ because that is the sample size used to make our CLT predictions for the mean and standard error of the mean. Because the samples must be random, you must be sure to sample with replacement. Thus, you would choose one score at random, put it back in the population, and then choose again at random.

15. All possible random samples with $N = 2$ are listed below (labeled A–P). These 16 samples are the population of sample means. Together, they are the distribution of sample means for the population of four billionaires when $N = 2$. Some of the means are computed for you. Complete the following table by finding the mean for the remaining three samples.

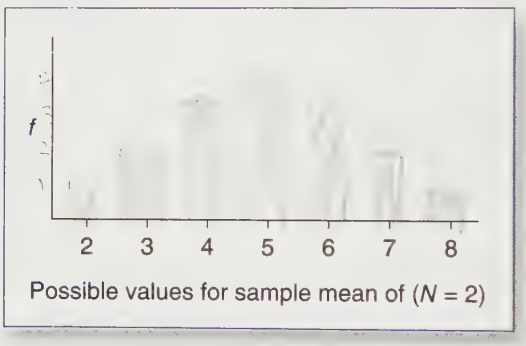
Sample	*First Score*	*Second Score*	*Sample Mean*
A	2	2	2
B	2	4	3
C	2	6	4
D	2	8	5
E	4	2	3
F	4	4	4
G	4	6	5
H	4	8	6
I	6	2	4
J	6	4	5
K	6	6	6

(Continued)

(Continued)

Sample	First Score	Second Score	Sample Mean
L	6	8	7
M	8	2	5
N	8	4	6
O	8	6	7
P	8	8	

The means you computed above are all the means that are possible when researchers take a sample of two scores from the preceding population of four people. Collectively, the means are the *distribution of sample means*. Fill in the following frequency distribution graph with all the possible sample means for this study when $N = 2$.



16. You should recognize that some of these samples represent the population better than others and therefore some have less *sampling error* than others. Each of the above sample means that are not exactly equal to the population mean of 5 ($\mu = 5$) have *sampling error*. Which two samples have the most sampling error?

    a. Samples A and P
    b. Samples B and O
    c. Samples G and H

17. You should also recognize that all of the above sample means are possible when the researcher randomly selects a sample from the population. Looking at the distribution of sample means, imagine that you randomly pick one sample mean from all the possible sample means. Which sample mean are you most likely to pick?

18. How does the graph of the distribution of sample means (the frequency distribution you created in Question 15) *differ* from the graph of the original data (the frequency distribution under Question 14)? (Choose two.)

a. Central tendency (i.e., mean)
b. Spread (i.e., variability)
c. Shape

19. Although the graphs look very different, there is one similarity between the graphs. Which of the following is identical for the two graphs?
a. Central tendency (i.e., mean)
b. Spread (i.e., variability)
c. Shape

20. Compute the mean of the distribution of sample means (i.e., compute the mean of the 16 sample means in the last column of the table under Question 15). You should be able to use the statistics mode on your calculator to obtain the mean.

21. Compute the standard deviation of the distribution of sample means. You should be able to use the statistics mode on your calculator to obtain the standard deviation. Because you have the population of all possible sample means, you should use the population equation for the standard deviation. If you choose to do the calculations by hand, be careful about what N you use. When computing the standard deviation of all 16 sample means, N is the number of sample means (i.e., 16).

22. How does the mean of the distribution of sample means compare with the mean of the population?
a. Same
b. Higher
c. Lower

23. How does the standard deviation of the distribution of sample means compare with the standard deviation of the population?
a. Same
b. Higher
c. Lower

24. Sample means are _____ variable than individual scores; therefore, the standard deviation of the distribution of sample means is _____ than the standard deviation of the population.
a. less, larger
b. less, smaller
c. more, larger
d. more, smaller

25. How did the CLT's predictions for the mean and standard error of the mean compare to those you found by actually building the distribution of sample means and then computing its mean and its standard deviation? Were the predictions made by the CLT perfectly accurate? (Hint: Compare the answers for Questions 20 and 21 to those predicted for Questions 12 and 13, respectively.)

   a. Yes, the answers were identical.

   b. The CLT predictions were close to the actual data but not identical.

DISTRIBUTIONS OF SAMPLE MEANS WITH LARGER POPULATIONS AND SAMPLE SIZES

In the previous questions, you worked with a very small population and sample size and learned that distributions of sample means are normally distributed and have a mean equal to the population mean (μ) and a standard deviation equal to the standard error of the mean (SEM_p).

26. Next, we are going to use a Java Applet to help you understand sampling distributions with larger (i.e., more realistic) populations and sample sizes. To access the Applet, go to this website: http://onlinestatbook.com/stat_sim/sampling_dist/index.html. When you get to this page, you will need to click on the "Begin" button. At the top of the screen, there is a "parent" population of scores. What is the mean and standard deviation for this population distribution of scores (they are located on the left side of the graph)?

   Mean (μ) = 16; Standard deviation (σ): 5

27. It might help to have a bit of context while you are looking at these distributions, so assume that these numbers represent scores on a test of extraversion, with higher numbers indicating greater extraversion. Suppose that you are going to do a study to determine if extraversion is related to levels of depression. Before starting the study, you want to make sure that you have a good sample with minimal sampling error. What is sampling error?

   a. The difference between scores and the sample mean

   b. The difference between scores and the population mean

   c. The difference between a sample mean and the population mean

28. Suppose that you plan to take one random sample with $N = 5$ from a population with $\mu = 16$ and $\sigma = 5$. According to the CLT, what is your best guess for the mean of this one sample?

    16

29. According to the CLT, how much sampling error do you expect to have (i.e., what value will the standard error of the mean have)?

30. You can use the Applet to simulate taking one sample at a time from the population. To do this, go to the graph that is labeled "Distribution of means" and select an $N = 5$. Then, click on the box that says "Animated." This will randomly select five scores from the population and put them in the second graph labeled "Sample Data." The mean of those five scores will then be graphed on the third graph labeled "Distribution of means." After you have taken one sample of five people, what is in the second graph labeled "Sample Data"?

    a. The sample mean based on the five scores

    b. The five scores selected randomly from the population

31. After you have taken one sample of five, what is in the third graph labeled "Distribution of means"?

    a. The sample mean of the five scores

    b. The five scores selected randomly from the population

32. If you click on the "Animated" button again, the program will randomly select another five scores and then put their mean on the Distribution of Sample Means graph. After doing this, how many sample means are on that graph?

    a. 1

    b. 2

    c. 5

33. Click on the Animated button several times and make sure that you understand what is happening when you hit the Animated button. After you have several sample means on the Sample Means graph, use the graph to determine which samples have the most sampling error. Remember, sampling error is the discrepancy between a sample statistic and a population parameter. How can you tell which samples have the most sampling error?

    a. The samples closest to the population mean ($\mu = 16$) have the most sampling error.

    b. The samples furthest from the population mean ($\mu = 16$) have the most sampling error.

34. The study of extraversion will use just one of the possible random samples of $N = 5$ people. We can't know ahead of time which sample mean that will be. However, the central limit theorem allows us to determine what the distribution of all possible random samples with $N = 5$ would look like. Then, we can determine the most likely sample mean and expected sampling error. If you take all possible random samples of a given size ($N = 5$) from the population with a mean of 16 and a standard deviation of 5, what do you expect the mean of the distribution of sample means to equal?

35. What do you expect the standard deviation of the distribution of sample means to equal?

    a. 5

    b. $\frac{5}{\sqrt{16}}$

    c. $\frac{5}{\sqrt{5}}$

36. What shape do you expect for the distribution of sample means?
    a. Normal, bell shaped
    b. The same shape as the original distribution of scores

37. To create a distribution of sample means, we need to obtain all possible random samples of a given size from this population. This program doesn't really allow us to take all possible random samples, but it does allow us to take a very large number of samples from the population, and this very large number is a very good approximation of what happens if you take all possible random samples. To take 100,000 samples at a time, click on the "100,000" button. You can even click it 10 times to take 1,000,000 samples. After you do this, what are the mean and the standard deviation of the distribution of sample means?

    Mean = 10 ; Standard deviation = _____

38. How close are the values to what the central limit theorem predicted in Questions 33 and 34?
    a. Very close (i.e., within .01)
    b. Not very close (i.e., not within .01)

39. The *population* standard deviation of $\sigma = 5$ tells us that the typical distance the individual scores are away from the population mean ($\mu = 16$) is 5. What does the standard deviation of the distribution of sample means (*SEM* = 2.24) tell us?
    a. The typical distance between sample means is 2.24.
    b. The typical distance sample means are from the population mean is 2.24.

40. How could we make the standard error (*SEM*) smaller? Select all that apply.
    a. Increase *N*
    b. Decrease *N*
    c. Increase σ
    d. Decrease σ

41. Why would we want to make the standard error smaller?
    a. It will reduce sampling error.
    b. The sample mean will be more likely to be close to the population mean.
    c. Both of the above

42. In the previous problems, you were planning to use a sample size of 5. What would happen if you used a larger sample (e.g., 25)? Use the CLT to compute the expected amount of sampling error with a sample size of 25 for the population with a mean of $\mu = 16$ and a standard deviation of $\sigma = 5$.

43. Click on the "Clear lower 3" button to clear the graphs. Create a distribution of sample means changing the N to $N = 25$ and then clicking on the "100,000" button. How close is the standard deviation of the distribution of sample means to the one predicted by the CLT (i.e., the one you just computed)?

    a. Very close (i.e., within .01)

    b. Not very close (i.e., not within .01)

44. How is the graph of the distribution of sample means with this larger sample size ($N = 25$) different from the distribution of sample means when the sample size was smaller ($N = 5$)?

    a. It is less variable (sample means are closer to the population mean).

    b. It is more variable (sample means are farther from the population mean).

45. As the variability in the distribution of sample means decreases, the amount of sampling error _____.

    a. increases

    b. decreases

    c. stays the same

46. In all the preceding examples, you started with a normally distributed population of extraversion scores. You are also interested in depression scores. The measure of depression is not normally distributed but skewed with a mean of $\mu = 8.08$ and a standard deviation of $\sigma = 6.22$. According to the central limit theorem, what would you expect for the mean and standard deviation of the distribution of sample means for samples of a size of 25 from a positively skewed population with $\mu = 8.08$ and $\sigma = 6.22$?

    Mean = 8.08

    Standard deviation of the distribution of sample means (*SEM*) = 1.244

47. According to the central limit theorem, what shape will the distribution of sample means start to approach if all possible random samples with an $N = 25$ are taken from a positively skewed population with $\mu = 8.08$ and $\sigma = 6.22$?

    a. Positively skewed

    b. Normally distributed

48. Use the Java Applet to see what happens for skewed distributions of scores. Start by changing the first graph to "Skewed." Then, select 100,000 random samples with a size of 2. Do the same with all of the sample size options (i.e., 5, 10, 16, 20, 25). As the sample size increased, what happened to the shape of the distribution of sample means?

    a. It stayed positively skewed.

    b. It became more bell shaped.

49. The Java Applet also allows you to change the shape of the distribution to anything you want. Make a really strange-looking distribution of scores by holding the mouse button and dragging it

over the population of scores. The Applet will update the mean and standard deviation of those scores. Record them below:

Mean (μ) = 6.06; Standard deviation (σ) = _____

50. According to the central limit theorem, what do you expect the mean and the standard deviation of the distribution of sample means to be for all possible random samples of 25 from the "strangely" shaped distribution you created?

Mean of the distribution of sample means = _____

Standard deviation of the distribution of sample means (*SEM*) = _____

51. According to the central limit theorem, what shape do you expect the distribution of sample means to approach when using all possible random samples of a size of 25?

a. Shaped like the parent distribution (strange looking)

b. Normally distributed (bell shaped)

52. Take 100,000 random samples of a size of $N = 25$ from the strange-looking distribution of scores you created. What does the distribution of sample means look like?

a. Shaped like the parent distribution (strange looking)

b. Approximately normally distributed (bell shaped)

53. Even if a parent population's distribution is skewed, the distribution of the sample means is approximately _____ and becomes even more so as the sample size _____.

54. What shape of a distribution allows you to use the unit normal table (i.e., the *z* table) to determine the probabilities associated with different *z* scores?

a. Normal (bell shaped)

b. You can use the unit normal table with any shaped distribution.

Activity 5.2: Central Limit Theorem

Learning Objectives

After reading the chapter and completing the homework, you should be able to do the following:

- Determine when to use the two different *z* score formulas
- Define sampling error and the standard error of the mean
- Compute and interpret the standard error of the mean
- Explain the difference between raw scores and *z* scores
- Compute and interpret the *z* score for a sample mean

DESCRIBING SAMPLING ERROR

1. In words, sampling error is the difference between a population parameter and a sample statistic?

2. In words, what is the standard error of the mean? Select all that apply.
   a. The standard deviation of the scores in a population
   b. The typical distance between scores and a population mean
   c. The standard deviation of the distribution of sample means
   d. The typical distance between sample means of a given size from a population mean

3. What is the relationship between sampling error and the standard error of the mean?
   a. You compute the standard error when you have a population and the sampling error when you have a sample.
   b. The standard error of the mean is a measure of sampling error.
   c. You divide the standard error of the mean by N to get a measure of sampling error.

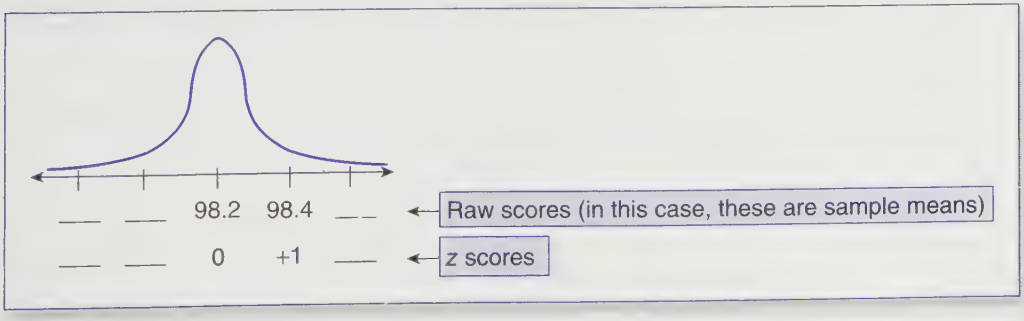
COMPUTING SAMPLING ERROR

The next several questions are working with body temperature data in which the population mean (μ) is 98.2 and the population standard deviation (σ) is 0.6.

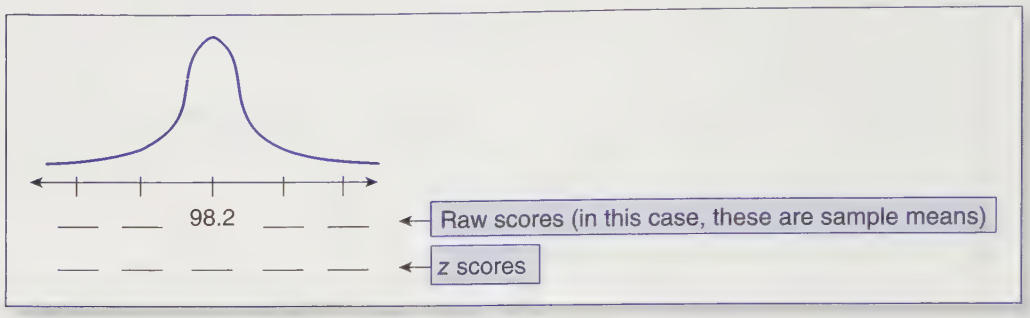
Compute the standard error of the mean for each of the following situations and use it to help you complete the correct values for the x-axes below. The standard error of the mean is equal to the standard deviation divided by the square root of N. Therefore, when $N = 9$, the standard error of the mean is 0.2 $\left(SEM_p = \frac{\sigma}{\sqrt{N}} = \frac{.6}{\sqrt{9}} = 0.2 \right)$. Thus, the standard deviation of the distribution of sample means is 0.2. You should also complete the number lines by filling in any values missing from the x-axes. The first one is started for you.

Each raw score can also be converted into z scores. For example, a sample mean of 98.4 can be converted into a z by using the z for a sample mean formula $= \frac{M - \mu}{SEM_p} = \frac{98.4 - 98.2}{.2} = 1$. You should also complete the number lines by filling in any missing z scores.

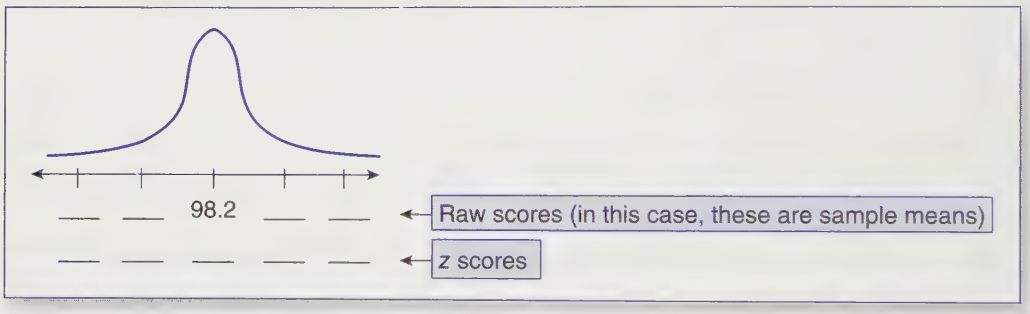
4. Distribution of sample means when sample size (N) is nine people: $\mu = 98.2$, $\sigma = 0.6$, $N = 9$.



5. Distribution of sample means when sample size (N) is 36 people: $\mu = 98.2$, $\sigma = .6$, $N = 36$.



6. Distribution of sample means when sample size (N) is 100 people: $\mu = 98.2$, $\sigma = 0.6$, $N = 100$.



7. Each of the examples above comes from the same population with $\mu = 98.2$ and $\sigma = 0.6$. Because of the way we have drawn the graphs, the distributions may look similar. However, there is one important difference between these three graphs that is created by the different sample sizes. If you look at the range of possible sample means on the x-axes of the graphs, you will notice that they are dramatically different. Which of the following best summarizes the differences between the graphs?
   a. When the sample size is larger, the range of values on the x-axes is smaller (i.e., the standard error is smaller).
   b. When the sample size is larger, the range of values on the x-axes is larger (i.e., the standard error is larger).

8. If all the graphs above were created using the same range of values on the x-axis (e.g., 97.8–98.6), which sample size would produce the most "spread-out" distribution of sample means? In other words, which sample size produces the most sampling error?
   a. 9
   b. 36
   c. 100

9. What is the probability of obtaining a *z* for a sample mean between −1 and +1? Note that this answer will be the same regardless of sample size. (Use the unit normal table to answer this question; there is an explanation of how to do this at the end of Chapter 4, right before the last reading question.)

10. A different sample size was used in each of the examples above. The size of the sample directly affects the probability of obtaining a sample mean close to the population mean. For each of the sample sizes listed below, write the sample means that correspond to *z* scores of −1 (*first blank*) and +1 (*second blank*). Refer to the graphs in Questions 4, 5, and 6 to do this. After you complete each of the three sentences that follow, read all of them carefully before you proceed to the next question.

    a. When the sample size is 9, 68.26% of all possible sample means are between the body temperatures of 98 and 100.4.

    b. When the sample size is 36, 68.26% of all possible sample means are between the body temperatures of 97.1 and 98.3.

    c. When the sample size is 100, 68.26% of all possible sample means are between the body temperatures of 98.14 and 98.26.

11. As sample size increases, the sample mean you are likely to get will be closer to the population mean (μ).

    (a) True

    b. False

DETERMINING THE PROBABILITY OF A GIVEN SAMPLE

12. Researchers take a sample of 27 toddlers who were all born prematurely and score them on visual acuity. They want to know how this sample compares with the mean visual acuity score of the population of toddlers who were carried to full term: $\mu = 45$, $\sigma = 6$, $N = 27$, $M = 44$. What is the probability of obtaining a sample mean of 44 or lower?

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{44-45}{6/\sqrt{27}} = \frac{-1}{1.15} = -.87$$

.1922

13. Researchers commonly use a cutoff probability of .05 to determine if a sample mean is likely or unlikely. In other words, if a probability is less than .05, they consider the outcometo be "unlikely" to occur by chance. Is a sample mean of 44 that is basedon 27 toddlers likely or unlikely *to have come from the target population of full-term toddlers with* $\mu = 45$, $\sigma = 6$?

14. Researchers take a sample of 100 adult men from northwest Indiana and measure their cholesterol. They want to know how this sample compares with the national average cholesterol score for adult men: $\mu = 208$, $\sigma = 15$, $N = 100$, $M = 205$. What is the probability of obtaining a sample mean of 205 or lower? Is this sample mean that is based on 100 adult men likely or unlikely *to have come from a target population with* $\mu = 208$, $\sigma = 15$?

15. An exercise instructor takes a sample of 105 middle school students, all of whom are participants in a team sport, and measures each student's "athletic aptitude." The instructor wants to know if this sample's "athletic aptitude" is different from the national average athletic aptitude for middle school students: $\mu = 1{,}003$, $\sigma = 25$, $N = 105$, $M = 1{,}004.5$. What is the probability of obtaining a sample mean of 1,004.5 or higher? Is this sample mean that is based on 105 students likely or unlikely *to have come from a target population with* $\mu = 1{,}003$, $\sigma = 25$?

16. Why are the probabilities you computed in the examples above estimates rather than precise values?

    a. The samples are not truly random.

    b. The samples are too small.

PICKING THE CORRECT STATISTIC

For each of the following problems, choose if a researcher should use a z for an individual score or a z for a sample mean. Just choose which statistic is appropriate. Do not compute the z yet.

17. A school psychologist wants to identify students who may need extra help in their math course and so gives the students a standardized math test. The scores are normally distributed and have a mean of 500 and a standard deviation of 100. Students who score below 400 are given additional tutoring. *What percentage of students would you expect to receive scores at or below 400?* Choose the correct statistic.

    a. z for an individual score

    b. z for a sample mean

18. A car seat manufacturer develops a seat that is safest for people who are between 62 inches and 68.5 inches. If the average height of women in the United States is 64 inches with a standard deviation of 2 inches, *what percentage of women fall outside of the safest range?* Choose the correct statistic.

    a. z for an individual score

    b. z for a sample mean

19. A researcher knows that poverty is often associated with poor nutrition. The researcher wants to determine if poverty is also related to height. To test this association, he obtains a sample of 25 women who have incomes below the poverty line. Assuming that the average height of

women is 64 inches with a standard deviation of 2 inches, *what is the probability that sampling error would result in obtaining a sample mean that is below 63 inches for a sample of 25 women?* Choose the correct statistic.

a. z for an individual score b. z for a sample mean

20. The average worker in a manufacturing plant produces 27 units every hour with a standard deviation of 2.9. The manager of the plant took a sample of 25 workers and gave them special training designed to increase their productivity. After the special training was complete, the 25 workers produced 28.2 units every hour. *What is the probability of randomly sampling 25 workers from the general population and their mean productivity being 28.2 or higher?* Choose the correct statistic.

a. z for an individual score b. z for a sample mean

21. The manager of the manufacturing plant is responsible for distributing bonuses based on each worker's productivity. The mean productivity for all workers in the plant is 27 units per hour with a standard deviation of 2.9. The manager wants to give every worker in the top third a $200 bonus and every worker in the middle third a $100 bonus. The workers in the bottom third will get no bonus. What is the number of units produced per hour that separates those in the top third from those in the middle third? What is the number of units produced per hour that separates those in the middle third from those in the bottom third? (Hint: To do this problem, you first need to find the z scores with 1/3 (.333) of the scores in the tail. Then, you need to solve for X.) Choose the correct statistic.

a. z for an individual score b. z for a sample mean

22. Once you are sure that you have chosen the correct statistic, do the computations for Questions 17 to 21.

Chapter 5 Practice Test

1. In Chapter 1, we defined sampling error as the discrepancy between a sample statistic and a population parameter. How can you compute the expected amount of sampling error between the sample mean of a given size and the population mean?
   a. $\frac{\sigma}{\sqrt{N}}$
   b. μ
   c. σ^2
   d. You cannot compute this value. Sampling error is a conceptual definition, not a calculation.

2. Which of the following changes would reduce the amount of sampling error in a study?
   a. Increase the sample size
   b. Decrease the sample size
   c. Use data from a distribution of scores that is normally distributed
   d. Use data from a distribution of scores that is not normally distributed (i.e., positively or negatively skewed)

3. Scores on a test of mathematical ability have a mean of $\mu = 152$ and a standard deviation of $\sigma = 19.2$. What is the standard error of the mean for samples of a size of 50 taken from this population?
   a. .384
   b. 19.2
   c. 2.72
   d. 7.37
   e. 152

4. Scores on a test of mathematical ability have a mean of $\mu = 152$ and a standard deviation of $\sigma = 19.2$. What is mean of the distribution of sample means for all samples of $N = 25$ taken from the population?
   a. .384
   b. 19.2
   c. 2.72
   d. 7.37
   e. 152

5. Which of the following is a measure of the typical distance between sample means and a population mean?
   a. σ
   b. $\frac{\sigma}{\sqrt{N}}$
   c. μ
   d. σ^2

6. Which of the following is a measure of the typical distance between scores and a population mean?
   a. σ
   b. $\frac{\sigma}{\sqrt{N}}$
   c. μ
   d. σ^2

7. Scores on measure of science knowledge are positively skewed with a mean of 74 and a standard deviation of 8. A researcher wants to conduct a study with 90 people from the population and wants to know what the distribution of sample means will look like for that sample taken from the population. Which of the following best describes the distribution of sample means?
   a. It will be positively distributed with a mean of 74 and a standard deviation of 8.
   b. It will be positively distributed with a mean of 74 and a standard deviation of .84.
   c. It will be normally distributed with a mean of 74 and a standard deviation of .84.
   d. It will be positively distributed with a mean of 74 and a standard deviation of .48.

8. Scores on measure of science knowledge are positively skewed with a mean of 74 and a standard deviation of 8. A researcher wants to conduct a study with 90 people from the population. Which of the following best describes what that distribution of 90 scores will most likely look like?
   a. It will be positively skewed.
   b. It will be bell shaped (normal).

9. The average salary for Lexar Steel Corporation is $52,600 with a standard deviation of $\sigma = 16{,}800$. A researcher obtains a random sample of 40 workers. What is the probability of obtaining a sample of workers with a mean salary greater than $60,000?
   a. .9974
   b. .3300
   c. .6700
   d. .0026
   e. .4400

10. The average salary for Lexar Steel Corporation is $52,600 with a standard deviation of $\sigma = 16{,}800$. What is the probability of one randomly selected worker having a salary greater than $60,000?
    a. .4400
    b. .3300
    c. .6700
    d. .0026
    e. .9974

11. The average grade point average (GPA) at a college is 2.9 with a standard deviation of .65. The dean of the college takes a random sample of 100 students. Which of the following is the most likely average GPA for that sample of 100 students?

    a. 3.2

    b. 3.5

    c. 2.8

    d. 2.4

12. A student wants to measure satisfaction with food in the college dormitories. To assess satisfaction, she plans to obtain a random sample of 35 students from the student directory and ask them all to answer just one question: How satisfied are you with the food choices in the dormitories? Responses are made on a 10-point scale where 1 = *very dissatisfied* and 10 = *very satisfied*. The student wants to present her data to the college administrators and wants to make sure that she will obtain an accurate estimate. What can she do to minimize sampling error?

    a. Nothing. This study is well designed. An N of 30 is what the central limit theorem requires and she plans to exceed that minimum.

    b. She should increase her sample size. A larger N would decrease sampling error.

    c. She should use a rating scale with more than 10 options. That will increase the variability in the sample and decrease the sampling error.

13. A population of test scores has a mean of 125 and a standard deviation of 5. Two separate researchers each take a sample from this population. Researcher A obtains a random sample of 25 people and Researcher B obtains a random sample of 100 people. Which researcher is more likely to obtain a sample mean of 135?

    a. Researcher A

    b. Researcher B

CHAPTER 6

Hypothesis Testing With *z* Scores

LEARNING OBJECTIVES

After reading this chapter, you should be able to do the following:

- Write null and research hypotheses using population parameters and words
- Compute a *z* for a sample mean
- Determine whether or not you should reject the null hypothesis
- Compute and interpret the effect size (*d*) of a study
- Summarize study results using American Psychological Association (APA) style
- Identify examples of Type I error, Type II error, and statistical power
- Provide a detailed description of a *p* value, critical value, and obtained value

INTRODUCTION TO HYPOTHESIS TESTING

In the previous chapter, you learned to (1) compute a *z* for a sample mean, (2) locate that sample mean within a distribution of sample means, and (3) determine the probability of that sample mean occurring due to sampling error. At this point, you may be wondering how these skills could be useful. The short answer is that all of these skills enable you to test hypotheses. Hypothesis testing is the most commonly used statistical procedure in psychology, sociology, social work, public policy, medicine, and many other fields. In fact, if your college major requires this course, it is probably because a working knowledge of hypothesis testing is considered critical by professionals in your discipline. Furthermore, hypothesis testing plays a central role in the scientific processes that permeate our society. Even if you do not plan to be a scientist, a working knowledge of how scientists use hypothesis testing is essential if you hope to understand contemporary issues like climate change or policy debates on health care, education, or voting rights, just to name a few. The hypothesis-testing process

that you are about to learn is powerful because it can be applied to such a wide range of topics. It can help all of us make better, more informed decisions, if we know how to use it.

HYPOTHESIS TESTING WITH *z* FOR A SAMPLE MEAN EXAMPLE (ONE-TAILED)

For example, suppose a teacher interested in improving educational practices reads a study reporting that frequent testing helps improve recall of information. She redesigns her history course so students are required to take frequent quizzes (two quizzes per chapter) prior to taking exams. After one semester using these quizzes, the mean score on the final exam for the 25 students in her course was $M = 80$. This same teacher is extremely organized and knows that the mean score on this same final exam for *all* students who took the course with her before she required frequent quizzing was 75 points with a standard deviation of 10 points ($\mu = 75$, $\sigma = 10$).

The teacher has every reason to believe that the 25 students in the new history course are similar to all the students she has taught in the past. So, logically, all the past students are considered a population and the 25 new students are considered a sample. Although these 25 new students were not literally randomly sampled from the past population, it is reasonable to expect them to score similarly to all previous students who took the same exam. With hypothesis testing, you can determine if the sample of students forced to take frequent quizzes scored "significantly" better than the previous students who did not take quizzes. If the quizzes helped the sample, perhaps requiring quizzes for all future students would help them as well. Did the frequent quizzing help? If the sample's mean final exam score is better than the population's mean final exam score, is the difference created by the quizzing or by sampling error?

Obviously, the sample mean of 80 is *numerically* greater than the population mean of 75. The teacher's question is whether the deviation between the sample mean of 80 and the population mean of 75 is likely or unlikely to have occurred due to sampling error. You can use the hypothesis-testing process and the *z* for a sample mean you learned in the previous chapter to determine if the 5-point difference is likely to be due to sampling error or if it is likely to have been created by the new frequent quizzing process. The computations necessary for this hypothesis test are identical to those you did in the previous chapter. However, we need to introduce quite a few concepts to evaluate hypotheses. These new concepts are described below in a series of six steps.

Step 1: Examine Variables to Assess Statistical Assumptions

All hypothesis tests are based on specific assumptions, and if these assumptions are violated, these tests *may* yield misleading results. Therefore, the first step when conducting any hypothesis test is to determine if the data you are about to analyze meet the necessary assumptions. Fortunately, if you understand four basic assumptions, you will be able to safely run all of the hypothesis-testing statistics discussed in this text. The four basic assumptions are (1) independence of data, (2) appropriate measurement of variables for the analysis, (3) normality of distributions, and (4) homogeneity of variance.

The independence of data assumption means that each participant's score within a condition is independent of all other participants' scores within that same condition. For example, in the study testing if frequent quizzing improves students' exam scores, you must ensure that all students answered every exam question on their own; if one or more students copied answers off another student instead of providing their own answers, this cheating would violate the assumption of data independence because one student's answers "influenced" another student's answers. If this cheating was widespread,

it would ruin the entire study. When assessing the independence of data, you are really assessing the procedural control used to collect the data. If a study's procedural control is very poor (e.g., there was widespread cheating or some other problem with data collection), the data violate the independence assumption, and you should redo the study more carefully. This is an important point; if a study's data collection procedures are very poor, it is not worth running a hypothesis test on the data because the results will be misleading. This is a research methodology issue, so we will not spend more time with this assumption in this book. You will learn how to design sound experiments that yield accurate data in a research methods course.

When conducting a hypothesis test using a z for a sample mean, the appropriate measurement of variables assumption means that the independent variable (IV) must identify a group of people who are different from the population in some way, and the dependent variable (DV) must be measured on an interval or ratio scale of measurement. If the IV or the DV does not meet these requirements, then the *z* for a sample mean is not the appropriate test. For example, in the frequent quizzing study, the IV is whether or not participants took frequent quizzes before they took the final exam. The IV identifies how the sample and population are different—namely, the sample took frequent quizzes and the population did not. When conducting a *z* for a sample mean, the IV is a "grouping variable"; it identifies a group of people who are different from the population. The DV in this study, students' final exam scores, must be measured so that each student's performance can be precisely quantified. In other words, the DV must be measured on an interval or ratio scale of measurement. If the exam scores are recorded as the number of correct responses or percentage correct, this assumption would be met because we would know exactly how much better (or worse) each student did relative to other students. A student with 12 correct answers did exactly two questions better than a student with 10 correct. A student earning an 80% did exactly 5% better than a student earning a 75%. Both of these possible DVs, number correct or percent correct, meet the interval/ratio scale of measurement requirement. If, however, the DV was measured by letter grades (e.g., A, B, C, D, or F), we would not be able to run the *z* for a sample mean test because this DV, letter grades, is ordinal, not interval or ratio. With letter grades, we do not know precisely *how much better* a student who earned an A did relative to a student who earned a B. In this study, the researcher looked at grades as a percentage, and so the DV was on an interval/ratio scale.

The normality of distributions assumption means that the distribution of sample means for each condition must have a normal shape. As you hopefully recall from the previous chapter, the central limit theorem tells us that a distribution of sample means will have a normal shape if (a) the original population of scores has a normal shape *or* (b) the sample size used in the study is 30 or larger. So, strictly speaking, in this study on frequent quizzing, the teacher should either determine if the original population of all students' scores has a normal shape or increase the sample size from 25 to 30 to be relatively confident that the normality assumption is met. However, in practice, unless researchers know that a distribution of scores is very skewed, they typically proceed with sample sizes that are "close to" 30. Ultimately, the soundness of this decision is based on how far a distribution's shape deviates from normal. If it deviates only slightly from a normal shape, a sample size "close to" 30 is probably sufficient. However, if a population's shape deviates substantially from normal, a sample even larger than 30 might be needed. In the end, researchers have to "know their variables" and make sound judgments. For this study, the teacher looked at the distribution of all past students' scores, and its shape was normal. Therefore, the distribution of sample means will have a normal shape regardless of sample size.

When conducting a z for a sample mean, the homogeneity of variance assumption means that the variances in each condition of the study are similar. In the frequent quizzing study, this means that the variance of scores when students experience frequent quizzes needs to be similar to the variance of scores when students do not experience frequent quizzes. Researchers cannot check this assumption until after they collect data. If the standard deviation of the sample is double that of the population or vice versa, this assumption might have been violated. In this study, the sample standard deviation was 9.5, and the population standard deviation was 10. Thus, the homogeneity of variance assumption was met.

Now that we have determined that all of the assumptions are likely to have been meet, we can proceed to Step 2.

Reading Question

1. The assumption of independence means that
   a. scores within a condition are independent of each other.
   b. the sample is independent of the population.

Reading Question

2. To conduct a hypothesis test using the *z* for a sample mean, the dependent variable must be measured on a
   a. ordinal scale.
   b. interval/ratio scale
   c. either ordinal or interval/ratio scale

Reading Question

3. When can you be fairly confident that the normality assumption is met?
   a. If the sample size is greater than 30
   b. If the population of scores is normally distributed
   c. If either the sample size is greater than 30 or if the population of scores is normally distributed

Reading Question

4. When conducting a *z* for a sample mean, you can be fairly confident that the homogeneity of variance assumption has been met by the data in a study as long as
   a. the standard deviations for the sample and the population are similar (i.e., one is not double the size of the other).
   b. the standard deviation for the sample and the population are more than double the standard deviation of the other group.

Step 2: State the Null and Research Hypotheses Symbolically and Verbally

The second step in hypothesis testing is to set up the null and research hypotheses. In this example, the teacher predicts that frequent quizzing will *increase* test scores. Thus, her research hypothesis states that the students who take frequent quizzes will have a *higher* mean than the population of students who did not take frequent quizzes. The formal symbolic representation of the research hypothesis is always in terms of population parameters. In this case, it is "$\mu_{quiz} > 75$." The "μ_{quiz}" represents the mean test score of the population *if* they all took frequent quizzes before the test. So, the research hypothesis of "$\mu_{quiz} > 75$" is predicting that *if* the population took the quizzes, their mean test score would be greater than it is now, 75. Less formally, the research hypothesis predicts that the treatment of quizzing will work. Quizzes will raise test scores.

Although the research hypothesis predicts that quizzing will increase test scores, it is possible that quizzing will *not* work. This second possibility is called the *null hypothesis*, and it is always the *opposite of the research hypothesis*. In this case, the null hypothesis is that frequent quizzing

will either have no effect on test scores or will *decrease* them. The formal symbolic representation of the null hypothesis is $\mu_{quiz} \leq 75$. Note that the null hypothesis includes the equal sign while the research does not. The null hypothesis of "$\mu_{quiz} \leq 75$" is predicting that if the population took the quizzes, their mean test score would be *less than or equal to* what it is now, 75. Less formally, the null hypothesis predicts that the treatment of frequent quizzing will *not* work.

You should recognize that the null and research hypotheses are opposites and, therefore, mutually exclusive. They collectively encompass all possible outcomes. Only one of these hypotheses can be correct. The entire point of doing hypothesis testing is to determine which of these two hypotheses is most likely to be true. The symbolic and verbal representations of both hypotheses are presented in Table 6.1.

Table 6.1 Symbolic and Verbal Representations of the Null and Research Hypotheses for the *z* for a Sample Mean

Hypothesis Type	*Symbolic*	*Verbal*	*Difference Between Sample and Population Means Was Created by*
Research hypothesis	$\mu_{quiz} > 75$	The population of people who take frequent quizzes *will* have quiz scores higher than 75.	The treatment improving final exam scores
Null hypothesis	$\mu_{quiz} \leq 75$	The population of people who take frequent quizzes *will not* have quiz scores higher than 75.	Sampling error

Reading Question

5. The research hypothesis states that
   a. frequent quizzes will increase final exam scores.
   b. frequent quizzes will either have no effect on final exam scores or will decrease them.

Reading Question

6. The null hypothesis states that
   a. frequent quizzes will increase final exam scores or will decrease them.
   b. frequent quizzes will either have no effect on final exam scores or will decrease them.

Reading Question

7. The fact that the null and research hypotheses are mutually exclusive means that if the null is true,
   a. the research hypothesis must also be true.
   b. the research hypothesis can be true or false.
   c. the research hypothesis must be false.

Some students find the symbolic notation of the null and research hypotheses a bit confusing because we are trying to determine whether or not $\mu_{quiz} > 75$, and we already know that $M = 80$ and $\mu = 75$,

based on the information given in the problem. However, it is important to note the subscript "quiz" in $\mu_{quiz} > 75$. The subscript refers to those who take frequent quizzes before the test. We do not know what the mean test score would be if the entire population of students took quizzes before the test (i.e., we don't know μ_{quiz}). We know that the population mean is 75 when students take the test without previously taking quizzes. The research question is whether taking frequent quizzes would create a population mean that is greater than 75.

Therefore, the research hypothesis is predicting that if there were a population of students who took quizzes before the test, their population test mean (μ_{quiz}) would be greater than 75 (i.e., the current mean of those who did not take quizzes before the test). In contrast, the null hypothesis is predicting that if there were a population of students who took quizzes before the test, their population test mean (μ_{quiz}) would be equal to or less than 75 (i.e., the current mean of those who did not take quizzes before taking the test). In this scenario, the research hypothesis is essentially saying the 5-point increase in exam score from 75 to 80 was created by the frequent quizzing. In contrast, the null hypothesis is saying that this increase was created by sampling error; frequent quizzing will not lead to an increase in the entire population.

Reading Question

8. In this research situation, μ_{quiz} represents
   a. the population's mean test score.
   b. what the population's mean test score would be if all students took quizzes before taking the test.

Step 3: Define the Critical Regions

The third step in the hypothesis testing process is defining the critical region. The critical region is the collection of all *z* score values that are so rare *if the null is true* that if any one of these *z* scores occurs, it suggests that the null hypothesis is probably false. To understand the critical region, we need to return to the distribution of sample means.

Figure 6.1 A Distribution of Sample Means With a *z* Number Line (*Top*) and a Sample Mean Number Line (*Bottom*)

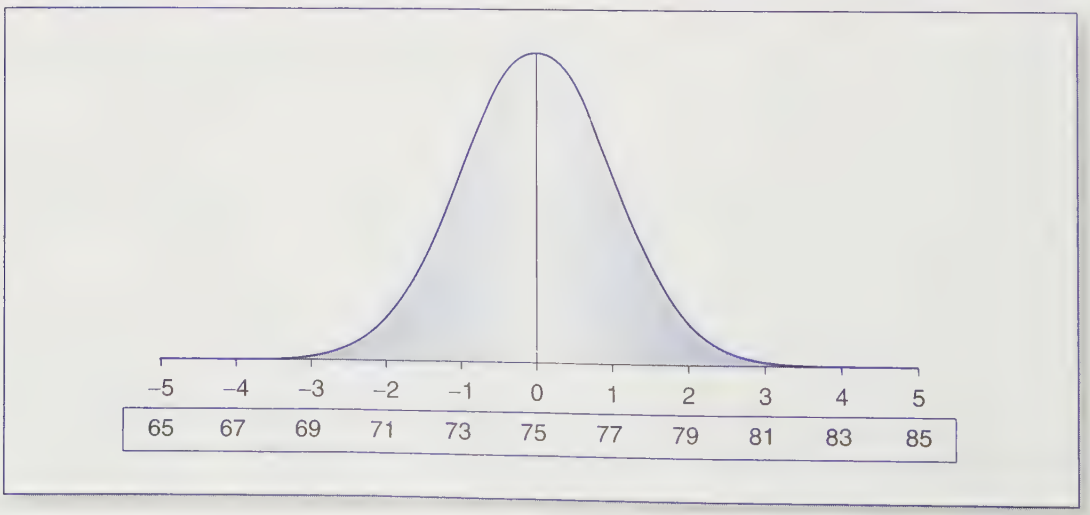


Figure 6.1 is the distribution of sample means for the frequent quiz study. It contains all possible sample means of size $N = 25$ taken from the population of students who took the final exam without completing frequent quizzes. The central limit theorem tells us that the mean of this distribution is 75 (μ), and that the standard error of the mean is 2 $\left(SEM = \frac{\sigma}{\sqrt{N}} = \frac{10}{\sqrt{25}} = 2 \right)$. Thus, the distribution is labeled with a mean of 75 and a standard deviation of 2. Ultimately, we will be working with z scores, so we have also converted each of the *possible* sample mean values on the x-axis into z scores using the z for a sample mean formula $\left(\text{e.g., } z = \frac{M - \mu}{SEM} = \frac{77 - 75}{2} = 1 \right)$.

If the null hypothesis is true (frequent quizzing has no effect on test performance), you would expect a z score of 0 because M and μ would be expected to be identical (i.e., they would both be expected to be 75). Therefore, the numerator of the z formula would be $M - \mu = 0$. But, even if the null hypothesis were true, the numerator of the z formula may not be exactly 0 because there is likely to be some sampling error. For example, because of sampling error, you wouldn't be surprised to get a z score of 0.3, even if quizzing has no effect on performance. You would, however, be surprised to get a z score of 2.9 if quizzing has no effect on performance.

Figure 6.1 is shaded to indicate the unlikeliness of specific z scores if the **null** hypothesis is true. The darker the shading, the more unlikely the z score is to occur if the null is true. The farther the z score is from zero (i.e., the darker the shading), the more likely it is that frequent quizzing had an effect on test scores.

Reading Question

9. If the null hypothesis is true, you should expect a z score that is close to
   a. 0.
   b. 1.
   c. −1.
   d. 2.9.

Reading Question

10. Even if the null hypothesis is true, you should not be surprised if the z score resulting from a significance test is not the exact value you selected above because of _____.
    a. a nonspecific research hypothesis
    b. sampling error

Reading Question

11. If the null hypothesis is true, z scores close to zero are _____.
    a. likely
    b. unlikely

While Figure 6.1 gives you a general idea of how probable particular z scores are if the null hypothesis is true, the unit normal table provides you with much more precise probabilities. For example, the probability of obtaining a sample mean with a z score of 1 or higher is .1587 (i.e., the tail area of $z = 1$) if the null hypothesis is true. The probability of obtaining a sample mean with a z score of 2.5 or higher is .0062 if the null hypothesis is true. Clearly, if the null is true, a z score of 2.5 is less likely than a z score of 1. However, hypothesis testing requires a cutoff that clearly identifies which z scores are unlikely if the null hypothesis is true. The most commonly used cutoff is a probability of .05. Thus,

z scores with probabilities of .05 or lower are considered unlikely to occur if the null hypothesis is true. z scores with probabilities less than .05 are unlikely to be due to sampling error, so they are more likely to be due to something else (e.g., frequent quizzing improving test scores). z scores with probabilities higher than .05 are considered relatively likely to occur if the null hypothesis is true; these z scores' deviation from 0 is likely to be due to sampling error.

It is important to remember that you could obtain a z score with a probability less than the .05 probability cutoff simply due to sampling error. In fact, *if the null were true,* you would have a 5% chance of getting an extreme z score, which would lead you to incorrectly conclude that the frequent quizzing improved test scores. This type of error is called a Type I error; we will explain this error in more detail later in this chapter. When you decide to use .05 as a cutoff *and the null hypothesis is actually true,* there is a 5% chance that you are going to get an extreme z score simply due to sampling error.

Reading Question

12. If you look up a z score of 2.2 in the unit normal table, you will find that the probability of obtaining a z score of 2.2 or higher if the null hypothesis is true is .0139. Is this a likely or unlikely outcome if we use a cutoff of .05?

    a. Likely

    b. Unlikely

Reading Question

13. If a researcher obtains a z score of 2.2, you can be absolutely confident that the observed difference between the sample mean and the population is not due to sampling error.

    a. True

    b. False

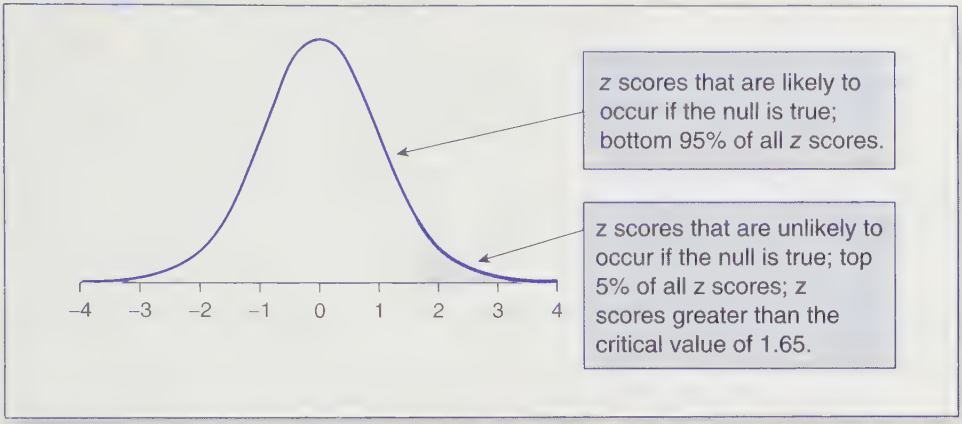
Figure 6.2 Finding The *Z* Critical Value That Defines the Critical Region When the Alpha Value Is .05

z score	Body	*Tail*
1.64	0.9495	*0.0505*
1.65	0.9505	*0.0495*
1.66	0.9515	*0.0485*

The *.05 cutoff value that is used to identify which z scores are unlikely if the null is true is called the* **alpha (α) level**. This alpha level cutoff can also be converted into a z score cutoff by looking up .05 in the tail column of the unit normal table and finding the z score with a tail of .05. Figure 6.2 illustrates how this is done. As you can see, the table does not include a tail probability of exactly .05, but there are two probabilities that are close to .05: .0505 and .0495. In Figure 6.2, we chose .0495 and used it to find the z score cutoff of 1.65. Forced to choose between .0505 and .0495, most statisticians would recommend using .0495 because it ensures that the Type I error rate is not more than 5% rather than slightly more than 5%. The take-home message is that the z score cutoff we will use for a one-tailed .05 significance test will be 1.65 in this course.

Figure 6.3 may help explain the relationship between the alpha value and the critical value. The alpha value of .05 is represented by the shaded area on the right tail of the curve in Figure 6.3. The shaded area is the top 5% of the curve. *The value located at the beginning of unlikely z score area (the shaded area) is called the* **critical value.** In this case, the critical value is the z score of 1.65. The critical value is the z score cutoff; z scores equal to or greater than the critical value

Figure 6.3 The Critical Value That Defines the Critical Region When the Alpha Value Is .05



are considered unlikely to occur if the null is true. In contrast, z scores that are closer to zero than the critical value (1.65) are considered likely to occur if the null hypothesis is true.

It is important to note that the research hypothesis for the frequent quizzing example predicts that scores will be higher when using frequent quizzing than when not using frequent quizzing. Because the research hypothesis predicted that the sample mean would be *greater* than the population mean, the critical value is *positive* (i.e., +1.65). However, if the research hypothesis had predicted a lower sample mean than the population mean, the critical value would be negative (i.e., −1.65), and the shaded area, called the critical region, would have been on the negative tail of the above distribution.

Reading Question

14. When the research hypotheses is that the sample mean will be lower than the population mean, the critical region will be on the _____ side of the distribution because they are expecting that the obtained z score will be _____.

a. positive, positive
b. negative, positive
c. negative, negative

Reading Question

15. The alpha (α) level determines the location of the

a. z score.
b. z score cutoff that starts the critical region (i.e., the critical value of z).

The critical value separates the distribution into two regions: (1) z scores that are likely to occur if the null hypothesis is true and (2) z scores that are unlikely to occur if the null hypothesis is true. For this example, z scores that are greater than 1.65 are unlikely to occur if the null hypothesis is true. Thus, if you obtain a z score that is in this critical region (i.e., greater than 1.65), you will reject the null

hypothesis. This critical region is also called the "region of rejection" because if the z score is in the critical region, you should reject the null hypothesis. If the z score is not in the critical region, you will not reject the null hypothesis. We will talk more about this decision in Step 4.

Reading Question

16. The critical value defines the "cutoff point" for
    a. determining if the null hypothesis should be rejected or not.
    b. determining if the research hypothesis should be rejected or not.

Reading Question

17. If the computed z score is in the critical region, you will
    a. reject the null hypothesis.
    b. not reject the null hypothesis.

Reading Question

18. If the computed z score is not in the critical region, you will
    a. reject the null hypothesis.
    b. not reject the null hypothesis.

Step 4: Compute the Test Statistic (z for a Sample Mean)

4a. Compute the Observed Deviation Between the Sample Mean and the Population Mean

Find the deviation between the sample mean and the population mean.

$$(M - \mu) = (80 - 75) = 5.$$

4b. Compute the Deviation Expected Due to Sampling Error

Compute the expected amount of sampling error given the sample size.

$$SEM_p = \frac{\sigma}{\sqrt{N}} = \frac{10}{\sqrt{25}} = 2.$$

In this case, the standard error of the mean is 2. This means that when $N = 25$, the typical distance that all possible sample means are from the population mean is 2. This is the amount of deviation we expect between any sample mean and the population mean due to sampling error (i.e., the difference expected by chance).

4c. Compute the z for a Sample Mean

Compute the obtained z score associated with the obtained sample mean.

$$z = \frac{M - \mu}{SEM_p} = \frac{80 - 75}{2} = 2.5.$$

The obtained z score value, also called the obtained value, associated with the sample mean of 80 is 2.5. The obtained z score of 2.5 is more extreme than the critical value of 1.65, so this z score is located in the region of rejection (i.e., the critical region). As a result, you can conclude that the z score of 2.5 is unlikely to be due to sampling error and that the unexpectedly high z score results from frequent quizzing *improving* test scores. You can also look up the probability of a z score of 2.5 in Appendix A. If the null is true, a z score of 2.5 has a probability of .0062; this probability value is called the *p* value. Small p *values* also tell you to reject the null. *Whenever the z score is in the critical region, the* p *value will be smaller than the alpha value.* Therefore, there are two ways to determine if you should reject the null hypothesis: (1) if the obtained z score is in the critical region or (2) if the p *value* is less than the alpha value. Both of these indicate that the probability of the z score being due to sampling error is small. To review the logic of this significance test again briefly, the obtained z score should have been close to 0 if the null was true. The z score was not close to 0 (close being defined as closer to 0 than the critical region starting at +1.65), and so the null hypothesis is probably false. Therefore, the researchers conclude that the research hypothesis is probably true. In other words, based on these results, it *seems* like the frequent quizzing *improved* test scores. However, when reporting these results, one should be cautious. Although unlikely, it is possible that the result occurred due to sampling error. Therefore, when reporting the results, one should say things like "the evidence suggests" that frequent quizzing improves test scores and avoid saying things like "these results prove" that frequent quizzing improves test scores.

Reading Question

19. If a z score is in the critical region, the null hypothesis is probably _____ and the research hypothesis is probably _____.
    a. false; true
    b. true; false

Reading Question

20. If a z score is beyond a critical value (farther into the tail of the distribution), the researcher should
    a. not reject the null hypothesis.
    b. reject the null hypothesis.

Step 5: Compute an Effect Size, and Describe It as Small, Medium, or Large

The purpose of hypothesis testing is to decide whether the difference between a sample mean and a population mean can be attributed to a treatment effect or if it is due to sampling error. If the null hypothesis is rejected, the researcher is saying that the observed difference is most likely due to a treatment effect and not sampling error. A related but different topic is the size of the difference between the sample mean and the population mean. For example, the researcher in this example concluded that frequent quizzing seems to improve test scores, but *how much* did frequent quizzing improve test scores? This question cannot be answered with hypothesis testing. Rather, this question requires another statistical procedure called an effect size.

An effect size is an index of the difference between the sample mean and the population mean. In this case, the effect size indicates how large the difference is between the mean for the sample that took frequent quizzes and the mean for the population that did not take frequent quizzes. The effect size statistic typically used when comparing two means is d. The numerator for the d is the same as the z for a sample mean (i.e., the difference between the means). The denominator, however, is different. When

computing the *z* for a sample mean, you divide by the standard error of the mean (i.e., typical sampling error). When computing *d*, you divide by the population standard deviation. The computations for this problem are shown as follows:

$$d = \frac{\text{Observed deviation between the means}}{\text{Standard deviation}} = \frac{M - \mu}{\sigma} = \frac{80 - 75}{10} = 0.50.$$

By computing *d*, you can describe the relative size of the deviation between the sample mean and the population mean. A $d = .5$ means that the sample mean was half a standard deviation higher than the population mean. Is a $d = .5$ good? The best way to interpret any effect size is by comparing it with the effect sizes produced by similar studies in the research literature. For example, if most studies investigating ways to improve test scores in the literature report effect sizes of $d = .25$ and your study had a $d = .5$, you should be pretty excited because your effect size is double that reported in the literature. Relative to other similar studies, your study created an impressively larger effect. However, it is not always possible to compare your effect size with that reported in the literature. If you can't find similar studies in the literature to provide a reference, you can use the general guidelines that Cohen (1992) suggested for interpreting effect sizes in Table 6.2.

Table 6.2 General Guidelines for Interpreting *d*

d	*Estimated Size of the Effect*
Close to .2	Small
Close to .5	Medium
Close to .8	Large

Cohen's *d* values close to .2 are small, those close to .5 are medium, and those close to .8 are large. This means that a *d* of .35 is a "small to medium effect." Again, these general guidelines are intended to aid interpretation when a larger context is not provided by studies in the relevant research literature.

In the present example, we do not have similar studies to provide specific guidelines for interpreting our *d* of .50. Therefore, we use general guidelines and conclude that the size of the difference between the quiz and no quiz means was medium. As stated earlier, a *d* of .50 indicates that the mean for those who took frequent quizzes is .50 standard deviations above the mean for the population, which did not take frequent quizzes.

Reading Question

21. An effect size of $d = .65$ is
    a. small.
    b. small to medium.
    c. medium.
    d. medium to large.
    e. large.

Reading Question

22. An effect size of .3 indicates that the means were different by
   a. .3 of a standard deviation.
   b. .3%.

In this example, we rejected the null hypothesis and found that the size of the effect quizzing had on the test scores was medium. We will address the important differences between hypothesis testing and effect size in an activity. One very important difference between the two statistical concepts is that hypothesis testing is heavily influenced by sample size and effect size is not.

Step 6: Interpreting the Results of the Hypothesis Test Using a *z* for a Sample Mean

For the rest of the semester, you will be computing statistical tests, interpreting the results, and then writing summary statements about the results of the tests. These summary statements need to be in a very specific format. The following is a summary statement about the above statistical test.

The exam scores of students who took frequent quizzes over the material ($M = 80$, $SD = 9.50$) were significantly higher than the scores of those who did not take frequent quizzes ($\mu = 75$, $\sigma = 10$), $z\ (N = 25) = 2.50$, $p = .0062$, $d = .50$.

In this case, we would write "$p = .0062$" because the probability of obtaining a z score of 2.5 or higher is .0062 (obtained from the unit normal table). Generally, you should give the exact p value if you know it. In some cases (e.g., the next chapter), you will not know it, and then you would write "$p < .05$" if you rejected the null hypothesis or "$p > .05$" if you did not reject the null hypothesis.

WHAT DOES IT MEAN TO DESCRIBE SOMETHING AS "STATISTICALLY SIGNIFICANT"?

When writing summary statements of statistical results, the phrases "significantly higher," "significantly lower," or "significantly different" have very specific meanings. In a statistical context, these phrases simply mean that the results were unlikely to be due to sampling error. Therefore, whenever you reject the null hypothesis (i.e., whenever the obtained z value computed earlier in Step 4 is in the critical region), you can describe the results as "statistically significant." It is important to understand that "statistically significant" is not a synonym for important. Rather, it simply means that the results were unlikely to be created by sampling error. Conversely, when you fail to reject the null, you can say things like, "There was *no* significant difference" when describing the results. Again, in a statistical context, this simply means that the obtained z value was not in the critical region. A result that was "not statistically significant" or "not significant" might be very important. For example, if you found that a very expensive treatment for a psychological disorder was "not significantly different" from a placebo treatment, that would be a very important finding.

Reading Question

23. When statisticians describe a result as "statistically significant" or simply as "significant," they are saying that (Select four answers)
   a. the null hypothesis was rejected.
   b. the null hypothesis was not rejected.

c. the obtained value was in the critical region.
d. the obtained value was not in the critical region.
e. the results were unlikely to be created by sampling error.
f. the results were likely to be created by sampling error.
g. the p value was less than the alpha value.
h. the p value was greater than the alpha value.

ERRORS IN HYPOTHESIS TESTING

Whenever you make a decision regarding a null hypothesis, there is a chance that you made an error. If you incorrectly rejected the null (i.e., you concluded that the treatment worked when it did not really work), you made a **Type I error.** If you did not reject the null (i.e., concluded that the treatment did not work but it actually did work), you made a **Type II error.** For each type of error, the decision made about the null hypothesis was wrong; after all, they are called errors. Stated succinctly, a **Type I error** occurs *when you incorrectly reject the null.* Therefore, a Type I error can only occur when you reject the null. A **Type II error** occurs *when you incorrectly fail to reject the null.* Therefore, a Type II error can only occur when you do not reject the null.

Reading Question

24. If you reject the null, there is some probability that you have made a
a. Type I error.
b. Type II error.

Reading Question

25. If you fail to reject the null, there is some probability that you have made a
a. Type I error.
b. Type II error.

Of course, your decision about the null might be correct. In hypothesis testing, there are two types of correct decisions. Researchers can either correctly reject the null or correctly fail to reject the null. A study that has a high probability of *correctly rejecting the null* is said to have high **statistical power.**

Reading Question

26. Statistical power refers to
a. correctly rejecting the null.
b. correctly failing to reject the null.

Table 6.3 may help you understand the relationship between specific decisions regarding the null hypothesis and the occurrence of Type I errors, Type II errors, and statistical power.

Unfortunately, you can never be certain if you made a correct or an incorrect decision. After all, if you knew the "true state" of the null, there would be no need to conduct a hypothesis test. However, you can estimate the probability that you made a Type I or Type II error. In addition, you can compute the

Table 6.3 Situations Defining Type I Error, Type II Error, and Statistical Power

		True State of Null	
		Null Is False	***Null Is True***
Statistical decision	**Reject null**	Correct decision; statistical power	Type I error
	Fail to reject null	Type II error	Correct decision

probability that you correctly rejected the null hypothesis (i.e., statistical power) because doing so, among other things, can help you assess the quality of your study. A study with low statistical power (i.e., a value far below .80) is deemed less reliable by researchers. We will discuss the relationships among Type I error, Type II error, and statistical power in this chapter's activities. For right now, your goal is simply to know the definitions of these terms.

Reading Question

27. If you did not reject the null, you can be certain that you made a Type II error.
    a. True
    b. False

An example may help you think about these terms. You conducted a study evaluating the impact of frequent quizzing on students' grades. Because you rejected the null, you know that there is some probability that you made a Type I error. In other words, you concluded that quizzing seemed to work, but it is possible that quizzing really did not improve exam scores and instead the result was due to sampling error. Of course, you might have correctly rejected the null (i.e., statistical power). There is *zero* probability that you made a Type II error because you must fail to reject null to make a Type II error. You rejected the null, so you certainly did not make a Type II error.

The language of hypothesis testing can be a bit tricky because of double negatives (e.g., "fail to reject the null"). We will explain why the double-negative language is commonly used later in the chapter. Table 6.4 summarizes a variety of correct ways to describe Type I error, Type II error, and statistical power.

Table 6.4 Various Correct Ways of Describing Type I Error, Type II Error, and Statistical Power

Statistical Term	***Focusing on Null***	***Focusing on Treatment***	***Focusing on Null (2)***
Type I error	Rejecting the null when you should not reject it	Saying the treatment works when it doesn't work	Rejecting a true null
Type II error	Failing to reject the null when you should reject it	Saying the treatment does not work when it does work	Not rejecting a false null
Statistical power	Rejecting the null when you should reject it	Saying the treatment works when it does work	Rejecting a false null

Reading Question

28. If a researcher concludes that a treatment did not work but it really did work, what type of error did she make?
    a. Type I
    b. Type II

Reading Question

29. Rejecting a false null hypothesis is called
    a. Type I error.
    b. Type II error.
    c. statistical power.

Reading Question

30. Concluding that a treatment works when it actually does work is called
    a. Type I error.
    b. Type II error.
    c. statistical power.

Reading Question

31. Rejecting the null hypothesis when you should not reject it is called
    a. Type I error.
    b. Type II error.
    c. statistical power.

HYPOTHESIS TESTING RULES

The process of hypothesis testing is nothing more than a formal set of rules that researchers use to determine if the null hypothesis (H_0) is likely or unlikely to be true. If the null hypothesis is true, the researchers expect the computed statistic to be close to a specific value. When performing a *z* for a sample mean, researchers expect the *z* score to be close to 0 *if the null hypothesis is true*. Every *z* score has some probability of occurring. *If the null hypothesis is true,* then *z* scores close to 0 have a high probability and *z* scores far from 0 have a low probability. If the probability of a *z* score is low enough, researchers reject the null hypothesis. The logic of this decision is that if a *z* score has a low probability *if the null is true* and that *z* score still happens, then the null hypothesis probably is *not* true. But how low does the probability of a *z* score have to be before the null hypothesis is rejected? It must be lower than the alpha value. In the previous example, we used an alpha value of .05. Researchers typically choose between an alpha value of either .05 (5%) or .01 (1%). In general, if you want to minimize Type I errors, choose an alpha of .01, and if you want to maximize statistical power, choose an alpha of .05. In either case, whenever the *p* value is less than the alpha level, the null hypothesis is unlikely to be true and you should reject it.

Reading Question

32. If the null hypothesis is true, you should expect the *z* score to be close to
    a. the population mean.
    b. 0.
    c. 0.05.

Reading Question

33. When you reject the null hypothesis, you are concluding that the null hypothesis is
    a. definitely not true.
    b. probably not true.

Reading Question

34. If the *z* score has a low probability if the null hypothesis is true, you should _____ the null hypothesis and conclude that the research hypothesis is probably _____.
    a. reject; true
    b. not reject; true
    c. reject; false
    d. not reject; false

Reading Question

35. How low does the probability of a *z* score have to be before you should reject the null hypothesis?
    a. Less than .95
    b. Less than .05

The previous paragraph describes the decision-making process that underlies hypothesis testing in a general way. The next paragraph describes the same hypothesis testing process with a specific example.

Researchers attempted to increase science test scores by tutoring students. If tutoring does not have any effect on the science scores, researchers expect an obtained (i.e., a computed) *z* value "close to 0." The researchers decided on a one-tailed alpha value of .05. They used this information to look up the critical *z* score (i.e., the critical value) in the unit normal table and found that it was +1.65. The obtained *z* value of +2.2 was greater than the critical value, so the researchers rejected the null hypothesis. Getting a *z* value of +2.2 or larger if the null were true would be "unlikely." In fact, you can determine precisely how unlikely. Use the unit normal table (Appendix A) to look up *the probability of getting a z value of +2.2 or larger if the null were true*. You should find that the probability or *p* value is .0139. This means that only 1.39% of all *z* values are larger than +2.2 if you assume the null hypothesis is true. The probability of a *z* score being +2.2 or larger is .0139, which is less than the alpha value of .05 or 5%; therefore, the researchers rejected the null hypothesis and concluded that the tutoring helped improve students' science scores.

Reading Question

36. A researcher is expecting the sample mean to be greater than the population mean and so does a one-tailed test with an alpha of .05. What is the critical value for this test?
    a. 0.05
    b. 1.65
    c. 2.22

Reading Question

37. How do you obtain the critical value?
    a. Look up .05 in the tail column of the unit normal table and find the *z* score on that row of the table.
    b. Look up .05 in the body column of the unit normal table and find the *z* score on that row of the table.

Reading Question

38. In this problem, what percentage of *z* scores are equal to or greater than the critical value?

a. 5%

b. 95%

Reading Question

39. To determine if the null hypothesis should be rejected, the researcher compared the computed *z* score with the critical value. The null hypothesis was rejected because the obtained (i.e., calculated) *z* score was _____ than the critical value.

a. less extreme

b. more extreme

Reading Question

40. In addition to rejecting or failing to reject the null, researchers can state the probability of obtaining a particular *z* score if the null hypothesis is true. This probability is obtained by looking up the value in the tail column of the unit normal table across from the _____.

a. alpha level

b. obtained (i.e., calculated) *z* score

c. critical value

WHAT IS A *p* VALUE?

A *p* value is the probability of getting an obtained value or a more extreme value assuming the null hypothesis is true. It is used by researchers to determine whether or not they should reject a null hypothesis.

Figure 6.4 may help you visualize how obtained *z* values, critical values, and *p* values are related.

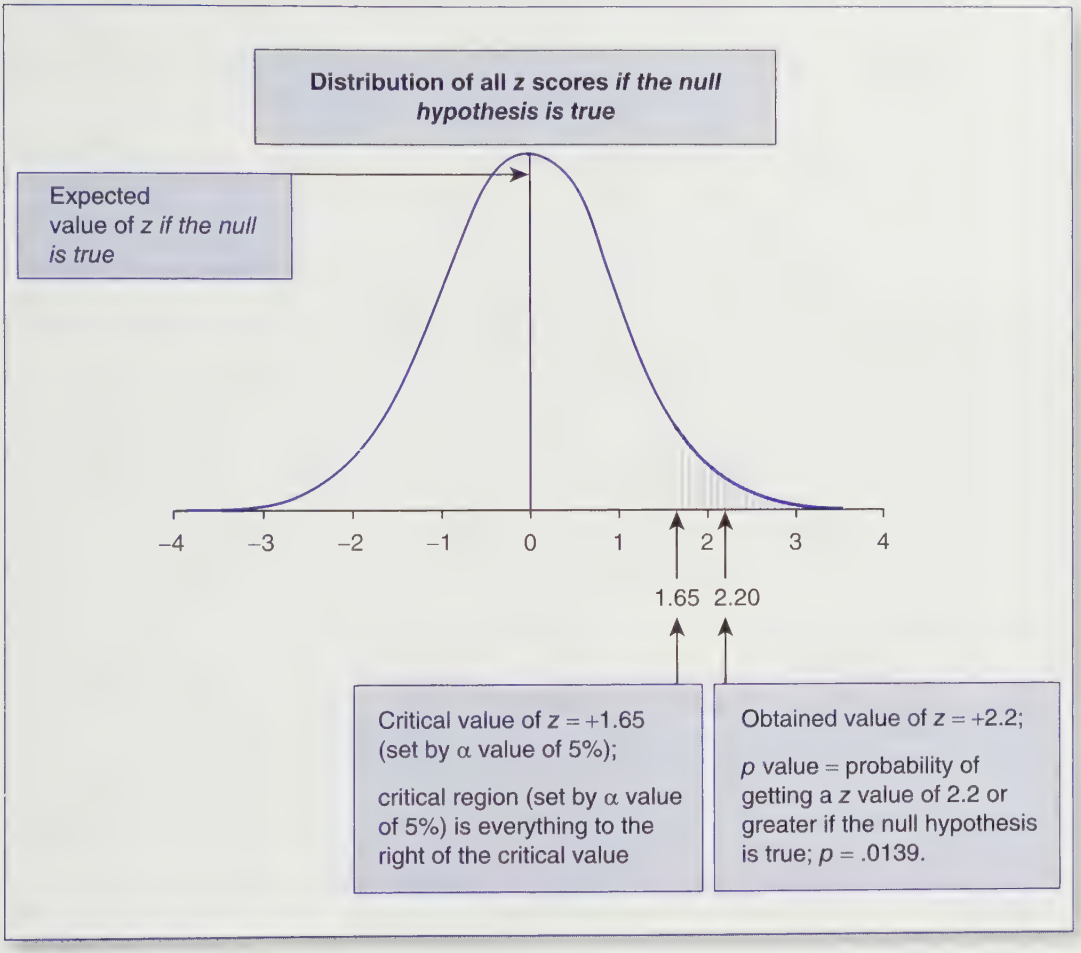
We will be performing several different statistical procedures in this book (e.g., *z*, *t*, *r*, *F*, and χ^2). For every one of these statistical tests, the basic rules of hypothesis testing are the same. *If the calculated statistical value (called the obtained value) is more extreme than the critical value, you should reject the null hypothesis.* In addition, *if the* p *value for an obtained value is less than or equal to the alpha value (usually .05 or .01), you should reject the null hypothesis.*

Whenever you are confronted with determining whether or not you should reject the null hypothesis, you can use *either* one of the following two rules:

1. If the obtained value is more extreme than the critical value, you should reject the null hypothesis.
2. If the *p* value is less than the alpha value, you should reject the null hypothesis.

Rule 1 will always lead to the same decision as Rule 2 and vice versa. These two rules are completely dependent. They are analogous to two different paths to the same location. It doesn't matter which of these two paths you take—you will always end up in the same place (i.e., the same statistical decision).

Figure 6.4 The Relationship Between Obtained z Values, Critical Values, and p Values



Reading Question

41. You should reject the null hypothesis if the p value is
   a. less than or equal to alpha ($p \leq \alpha$).
   b. greater than alpha ($p > \alpha$).

Reading Question

42. You should reject the null hypothesis if the obtained z value is _____ than the critical value.
   a. more extreme
   b. less extreme

WHY STATISTICIANS "FAIL TO REJECT THE NULL" RATHER THAN "ACCEPT THE NULL"

Students often wonder why statisticians use the confusing double negative "fail to reject the null" instead of the easier to understand "accept the null." We'll offer two related explanations. The first will be in the context of a specific research example. The second will be a more abstract logical argument.

A researcher testing the effectiveness of a program intended to reduce math anxiety administers the program to a sample of $N = 25$ people. After the participants complete the program, the researcher compares their mean anxiety score ($M = 55$) with the known population parameter ($\mu = 58$, $\sigma = 10$). The resulting z score for the sample mean of -1.5 is not farther from 0 than the critical value of -1.65, so he *fails to reject the null hypothesis*. The null states that the program is not effective. By failing to reject this hypothesis, the researcher is saying that the present study did not provide enough evidence to reject it. However, he is also acknowledging the possibility that the null hypothesis might not have been rejected because of one or more possible problems with the study. For example, a false null might not be rejected for a large number of reasons, including sampling error, if the program needed more time to be effective, or if there were any number of methodological problems with the study. By "failing to reject the null," the researcher is acknowledging all of these possibilities. So, the phrase "failing to reject the null" is saying that there is *currently* not enough evidence to reject the null hypothesis, but it acknowledges that future research might provide that evidence. In contrast, the phrase "accept the null" ignores the prospect that future research might change our conclusion. "Accepting the null" implies a final, definitive conclusion that belies most research situations. Ignoring alternative possibilities is not consistent with the cautious nature of scientific conclusions.

The second explanation for why researchers do not use the words "accept the null" is related to an inherent limitation of inductive reasoning. For example, suppose you have a hypothesis that all swans are white. You could test your hypothesis by going to a local pond. If you find that all swans at that pond are indeed white, does your visit to the local pond mean you should "accept" the hypothesis that *all swans* are white? No, it is possible that there is a different pond that does have at least one black swan. Even if you visit 1, 3, or 10 additional ponds and at every one you see only white swans, you cannot rule out the possibility that somewhere in the world there exists a pond with at least one black swan. In other words, you should not "accept" the hypothesis that all swans are white. Instead, you should only say that after many, many attempts to refute your hypothesis that all swans are white, you have "failed to reject" it every time. You might be very confident in your hypothesis, but scientifically you have to acknowledge that a black swan is possible. Similarly, when a researcher conducts many experiments (i.e., visits many ponds) and after every experiment he uses words like "failed to reject the null," he is acknowledging that another study could potentially reject that same null hypothesis.

Scientifically speaking, failing to reject the null hypothesis is a very uninformative outcome. If you take a research methods course, you will learn potential reasons why a study might fail to reject the null hypothesis. Only *one* of these reasons is that the null hypothesis is actually true.

Reading Question

43. Failing to reject the null hypothesis is the same as accepting the null hypothesis.
    a. True
    b. False

Reading Question

44. A researcher may fail to reject the null hypothesis because
    a. the treatment really did not work.
    b. of sampling error.
    c. the study was poorly designed.
    d. All of the above

Reading Question

45. Statisticians do not "accept" the null hypothesis because
    a. doing so fails to recognize the possibility of problems in their study.
    b. they don't not like speaking in double negatives.

WHY SCIENTISTS SAY "THIS RESEARCH SUGGESTS" RATHER THAN "THIS RESEARCH PROVES"

While we are on the topic of how to describe research conclusions, it is worth explaining why scientists use cautious, even tentative, language when describing research results. If you have listened to scientists talk on the radio or television or if you have read accurate quotes from scientists, you have probably heard them use the phrase "this research *suggests*" rather than the more impressive sounding "this research *proves*." Why do scientists intentionally avoid the word *proves*? The reason is closely related to the discussion in the previous section. For example, suppose a researcher conducted a study in which he rejected a null hypothesis and concluded that a tutoring program increases students' science scores on a standardized test. Based only on this result, if the scientist said, "This study proves that this tutoring program works," he would be ignoring the possibility that the study's result was due to sampling error (always a possibility). He would also be ignoring the possibility that the result was caused by some unknown flaw in the study's methodology. By using the more tentative, "this research suggests" language, the scientist is acknowledging that while the current information he has suggests the program works, it is possible that future work could change that conclusion. One of the reasons why the scientific process is so powerful is that it is willing to change its position if new evidence suggests that a new position is more accurate. If someone claims that he or she has the *final* answer, you can be sure that the person is not a scientist. Scientists always recognize that an answer can change as more research results accumulate.

So, does this mean that we can never be confident in any scientific conclusion? Well, no. If a given result has been replicated by many studies using different research methodologies, it is very unlikely that this conclusion would be reversed by future research. For example, many different research studies by many different research teams using many different methodologies have reported that frequent quizzing (i.e., testing) leads to improved memory scores (Rowland, 2014). This is such a common finding that it is extremely unlikely to be reversed by future research. While a single study cannot *prove* a conclusion, a multitude of studies all suggesting the same conclusion should make us *extremely* confident in that conclusion.

Reading Question

46. Statisticians do not say "this study proves" because
    a. there is always some probability that sampling error created the results.
    b. there is always a possibility a flaw in the methodology of the study created the results.
    c. Both of the above

OVERVIEW OF THE ACTIVITIES

In Activity 6.1, you will work with a specific research scenario to understand the steps of hypothesis testing. After you work through the steps, you will use the CLT to create distributions of sample means and use these distributions to understand Type I error, Type II error, and statistical power. Activity 6.2 is relatively short and is intended to serve as a review of the material covered in Activity 6.1. In addition, in Activity 6.2, you will work to understand the relationship between *p* values, alpha levels, and critical regions. In Activity 6.3, you will work with a Java Applet that will allow you to manipulate sample size, variability, treatment effects, and alpha levels to see their effect on Type I errors, Type II errors, and statistical power. Activity 6.4 reviews the previous material about hypothesis testing with the *z* for a sample mean and helps you distinguish between the purposes of hypothesis testing and computing effect sizes.

Activity 6.1: Hypothesis Testing

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Write null and research hypotheses using population parameters or words
- Create a distribution of sample means assuming the null hypothesis is true and a distribution of sample means assuming the research hypothesis is true
- Explain why the null hypothesis is necessary
- Define a critical region and use it to identify "likely" and "unlikely" sample means assuming the null hypothesis is true
- Compute a *z* for a sample mean and determine whether or not you should reject the null hypothesis
- Compute and interpret an effect size (*d*)
- Define statistical power, Type I error, and Type II error and locate them on a graph of the sampling distributions for the null and research hypotheses
- Determine if a given study had sufficient statistical power
- Explain why a larger sample size is usually preferable to a smaller sample size
- Distinguish among significance testing, effect size, and the practical usefulness of a treatment

AN EXAMPLE OF HYPOTHESIS TESTING (SIGNIFICANCE TESTING)

There are a number of steps involved in testing hypotheses. The mechanics of testing a hypothesis are not difficult, but the underlying logic requires quite a bit more explanation. In this activity, you will start by working through the steps of hypothesis testing to give you some experience with the process. Then, you will work through the steps for that same problem in far more detail. Finally, there is some

logic that is inherent in the process of hypothesis testing that is not obvious in the steps but is vitally important to your understanding. The last part of this activity explains that logic. Let's start with an example:

Suppose you are an educational researcher who wants to increase the science test scores of high school students. Based on tremendous amounts of previous research, you know that the national average test score for all senior high school students in the United States is 50 with a standard deviation of 20. In other words, *50 and 20 are the known population parameters for the mean and the standard deviation, respectively* (i.e., $\mu = 50$, $\sigma = 20$). You also know that this population of test scores has a normal shape. You and your research team take a sample of 16 high school seniors ($N = 16$), tutor them for 2 months, and then give them the national science test to determine if their test scores after tutoring were higher than the national average science test score of 50 (i.e., $\mu = 50$). After the tutoring, the mean test score of the 16 student sample was $M = 61$ ($SD = 21$). Now you need to determine if the difference between the sample's mean of 61 and the population mean of 50 is likely to be due to sampling error or if the tutoring improved the sample's science test score.

Because you are only interested in adopting the tutoring program if it *increases* science test scores, you use a one-tailed hypothesis test. You also chose a .05 alpha value (i.e., $\alpha = .05$) as the decision criterion for your significance test. If the sample mean has a chance probability less than .05, you will conclude that the tutoring program improved students' test scores.

Step 1: Examine the Statistical Assumptions

1. Your study meets all four of the necessary assumptions. Match each of the statistical assumptions listed below to the fact that suggests that assumption was met.

   ___ Independence

   ___ Appropriate measurement of the IV and the DV

   ___ Normality

   ___ Homogeneity of variance

   a. The population of science tests scores has a normal shape.

   b. The standard deviation of the sample was $SD = 21$ and the standard deviation of the population was $\sigma = 20$.

   c. All students in the study took the science test under controlled conditions, so each student's score reflects his or her own knowledge of science.

   d. The IV in this study is a grouping variable that identifies how the sample is distinct from the population, and the DV is measured on an interval/ratio scale of measurement.

Step 2: State the Null and Research Hypotheses

2. Write H_0 next to the symbolic notation for the null hypothesis and H_1 next to the research hypothesis.

   a. ___ $\mu_{tutoring} > 50$

   b. ___ $\mu_{tutoring} < 50$

   c. ___ $\mu_{tutoring} \geq 50$

   d. ___ $\mu_{tutoring} \leq 50$

e. ____ $\mu_{tutoring} > 61$

f. ____ $\mu_{tutoring} < 61$

g. ____ $\mu_{tutoring} \geq 61$

h. ____ $\mu_{tutoring} \leq 61$

3. Write H_0 next to the verbal description of the null hypothesis and H_1 next to the research hypothesis.

   a. ____ The population of students who receive tutoring will have a mean science test score that is equal to 50.

   b. ____ The population of students who receive tutoring will have a mean science test score that is greater than 50.

   c. ____ The population of students who receive tutoring will not have a mean science test score that is greater than 50.

   d. ____ The population of students who receive tutoring will have a mean science test score that is less than 50.

Step 3: Locate the Critical Region

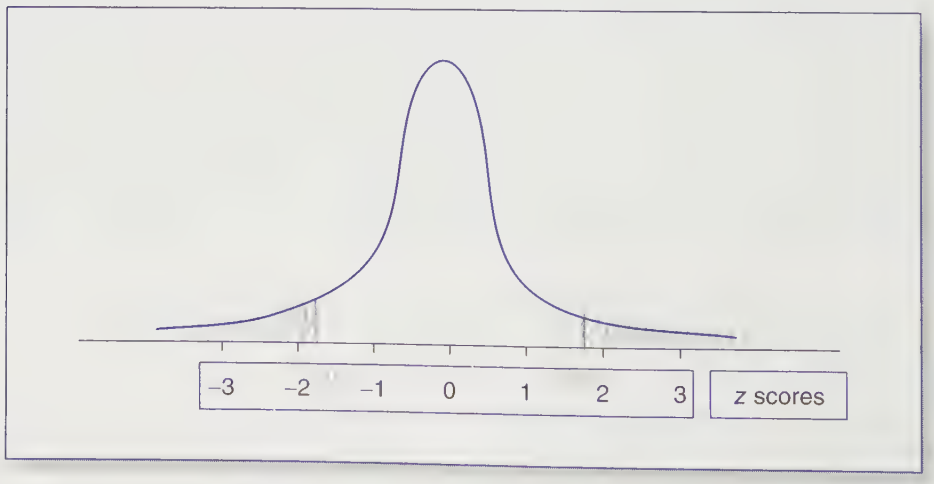
4. You chose to use a decision criterion of $\alpha = .05$. Consequently, the critical value for your hypothesis test is ____.

   a. 1.96

   b. 1.65

   c. .05

5. Draw a line at the critical value on the distribution of sample means drawn below. Then, shade the critical region. This region represents the sample means that are unlikely to occur if the null hypothesis is true and the treatment does not work.



Step 4: Compute the Test Statistic

6. Find the deviation between the sample mean and the population mean.

$(M - \mu) =$ 61 - 50 = 11

7. Compute the deviation expected due to sampling error.

$$SEM_p = \frac{\sigma}{\sqrt{N}} = \frac{20}{\sqrt{16}} = \frac{20}{4} = 5$$ SED = SD of sample

8. Compute the z for the sample mean.

$$z = \frac{M - \mu}{SEM_p} =$$ 11/5 = 2.2

9. What is the probability of obtaining a z score equal to or greater than the z score you computed in Question 8 (i.e., the p value)? $z \geq$ .0139

10. Should you reject or fail to reject the null hypothesis?

a. Reject. The z is in the critical region and the p is less than .05.

b. Fail to reject. The z is not in the critical region and the p is greater than .05.

11. Based on the statistical results, what should you say about the effect of tutoring on science test scores?

a. The results suggest that tutoring probably improved science scores.

b. The results suggest that tutoring probably did not improve science scores.

c. The results ~~prove~~ that tutoring improved science scores.

d. The results ~~prove~~ that tutoring did not improve science scores.

Step 5: Compute the Effect Size and Describe It

12. Compute d to determine how many standard deviations the scores improved by and describe how effective the tutoring treatment was.

$$d = \frac{M - \mu}{\sigma}.$$ $d = \frac{11}{20} = .55$

13. How effective was your tutoring program? Based on Cohen's recommendations, was this improvement small, small to medium, medium, medium to large, or large?

Step 6: Summarize the Results

14. Fill in the blanks of the APA style reporting statement:

The results suggest that the tutoring program improved science test scores. The sample that received the tutoring program had a significantly higher mean test score ($M =$ 61, $SD =$ 21) than in the untutored population's test score ($\mu =$ 50, $\sigma =$ 20), z ($N =$ 16) = 2.2, $p =$ .0139; $d =$ 5.5.

Being able to perform the steps of a hypothesis test is an important skill, but understanding the logic behind each of these steps is even more important. The next part of the activity uses the same example you completed earlier but focuses on the logic underlying each step of the hypothesis-testing process.

Understanding the Null and Research Hypotheses

15. The null and research hypotheses each make opposing and explicit statements about the entire *population* of senior students after they get tutoring. Of course, only a sample of the population actually received tutoring. You used the sample to infer what the entire population *would be like if it had received tutoring*. What is it called when you use a sample to represent a population and infer that the sample results represent what the population *would be like* if the entire population was treated like the sample was treated? Choose from the options provided below:

    a. Descriptive statistics
    b. Inferential statistics
    c. Sample statistics
    d. Population statistics

16. Whenever you use a sample to represent a population, your sample probably does not represent the population *perfectly*. Which of the following is created when a sample does not represent the population perfectly?

    a. Statistical error
    b. Parameter error
    c. Inferential error
    d. Sampling error

17. Assuming the null hypothesis is true (that tutoring does not work at all), what precise value would you *expect* from the *sample* of seniors that received tutoring? (Refer to the information before Question 1 if you need to.)

    Precise expected value for sample mean if null is true = 50.

18. Even if you assume that tutoring does not work at all, given that the study used a sample, you should not be surprised if the sample mean was not *exactly* the value you indicated above. Why not? Explain your answer. ? sampling error

19. The term for your explanation of Question 18 (and the answer to Question 16) is *sampling error*. In fact, you know exactly how to compute sampling error. Your study sampled 16 senior students. What is the value for expected sampling error in this scenario? (Hint 1: What is the formula for expected sampling error? Hint 2: The $\sigma = 20$.) SEM $\frac{\sigma}{\sqrt{N}}$ 5

20. Which of the following describes how you should interpret the sampling error value you computed above?

    a. When the sample size is $N = 16$, the typical distance between possible sample means and the population mean is 5 points.
    b. All possible means will be within 5 points above the population mean and 5 points below the population mean.

c. Every time a sample of 16 people is taken, it will be exactly 5 points away from the population mean.

d. A sampling error of 5 is large.

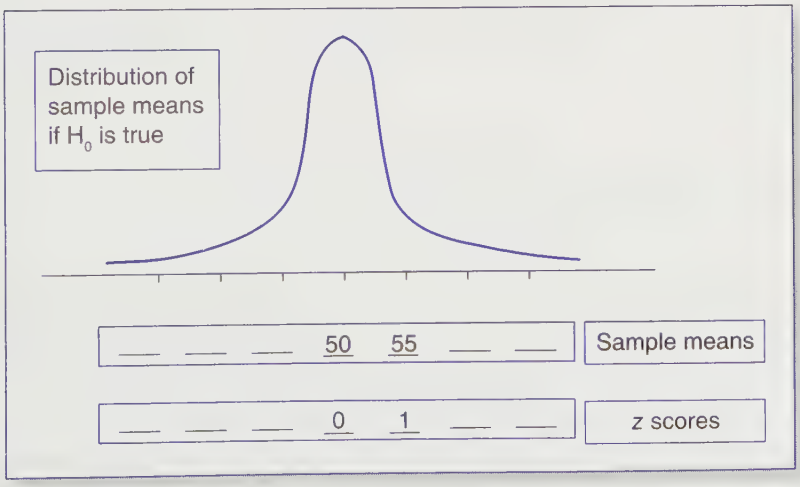
21. The central limit theorem tells you what the distribution of sample means should look like if the null hypothesis is true and tutoring does not work at all. The population of test scores had a normal shape and a mean of $\mu = 50$ with a standard deviation of $\sigma = 20$. You used a sample of 16 people. Based on this information, what is the shape, the mean, and the standard deviation (i.e., the standard error of the mean) of the distribution of sample means assuming the null hypothesis is true?

a. Shape: normal

b. Mean = 50

c. Standard error = 20

22. Use the curve provided below to complete the distribution of sample means that you would expect if the null hypothesis is true. First fill in the sample means. The center or peak of this distribution is located over the sample mean you expect if the null hypothesis is true (i.e., 50). One standard error to the right is 55. The values of 50 and 55 and their corresponding *z* scores are filled in for you. Complete the remaining sample means and *z* scores.



Defining the Critical Region

23. Draw a vertical line at the critical value of *z* and shade the critical region of *z*. You may need to consult the chapter to remind yourself how to locate the critical value of *z*.

24. Look at the distribution of sample means you just created. Researchers use the above distribution to determine if a sample mean is likely or unlikely to occur if the null hypothesis is true. Unlikely sample means cause researchers to reject the null hypothesis. The critical value of

z determines where the "cut line" between *likely* and *unlikely* sample means is located. Which of the following statements is most accurate?

a. The cut line's location is determined by the sample size.

b. The cut line's location is determined by the standard deviation of the population.

c. The cut line's location is determined by the population mean.

d. The cut line's location is determined by the alpha value.

25. The z score located exactly on the cut line is called the

a. absolute value.

b. critical value.

c. significant value.

d. immaculate value.

26. z Scores in the critical region are

a. unlikely if the null hypothesis is true.

b. likely if the null hypothesis is true.

27. Researchers reject the null hypothesis if the computed z score is

a. outside of the critical region.

b. inside the critical region.

28. If a z score is in the critical region with an alpha of .05, the p value for that z score must be

a. less than .05.

b. greater than .05.

Why Must We Assume the Null Hypothesis Is True? Why Do We Need a Null Hypothesis?

29. Go back to the second page of this exercise and reread the null and research hypotheses. (*No, really go back and read them!*) You will notice that the null hypothesis is very specific. In this case, when we assume that tutoring does not work at all, we can then *expect a very specific value* for the sample mean. What value does the null hypothesis say we should expect for the sample mean?

The null hypothesis tells us that we should expect the sample mean to be = _____.

30. In contrast to the very specific prediction of the null hypothesis, the research hypothesis is vague. In this case, if we assume that the one-tailed research hypothesis is true (i.e., that tutoring *increases* science test scores), we should expect the sample mean to be (choose one)

a. a specific value.

b. a value in a general range.

31. Given your answer to the previous two questions, why is it necessary to test the null hypothesis rather than the research hypothesis?

a. The null hypothesis provides a specific value to compare the sample mean to, whereas the research hypothesis does not provide a specific value to compare the sample mean to.

b. The research hypothesis provides a specific value to compare the sample mean to, whereas the null hypothesis does not provide a specific value to compare the sample mean to.

Type I Errors, Type II Errors, and Statistical Power

32. Thus far, you have only looked at the distribution of sample means assuming the null hypothesis is true. We started there because we know exactly what to expect if the null hypothesis is true. *Before* collecting data from the sample that received tutoring, you can't locate the center of the distribution of sample means if the *research hypothesis* is true. However, *after* you have the data from the sample, you do have *an idea* of where the research hypothesis distribution of sample means is located. What could you use to estimate the location of this distribution? Provide a specific value. (Hint: What value might represent what the population's science score would be if they all received tutoring?)

    Specific value for the center of the research hypothesis distribution of sample means = _____.

33. The research hypothesis distribution of sample means represents all possible sample means if the research hypothesis is true. For this one study, you have just one sample mean. Therefore, if a different sample of people participated in the study, it would likely produce a different sample mean. Suppose that another study was done with 16 different people and the sample mean was $M = 58$ rather than 61. How would the research hypothesis distribution of sample means for this second study differ from that of the first study?

    a. It would be more spread out than when the sample mean was 61.

    b. It would be centered on 58 rather than 61 and more spread out.

    c. It would be centered on 58 rather than 61, but its spread would be the same.

34. Your answer to Question 32 is considered an *estimate* of the mean of the research distribution of sample means. Why do we have to "guess" or estimate the mean of the research distribution of sample means?

    a. Sample means are more variable than population means, and so we cannot predict the exact value of the sample mean if the null hypothesis is true.

    b. Before we collect data, all we have is a prediction that scores will be higher than the population mean, but we don't know *how much* higher the sample mean will be than the population mean.

35. Even though the center of the research hypothesis distribution of sample means is based on an estimate, if the sample's size was sufficiently large, it can be thought of as a close estimate. Which of the following explains why a larger sample size would improve one's confidence in the estimate?

    a. A larger sample size increases the standard error of the mean.

    b. A larger sample size reduces the amount of sampling error we expect a sample mean to have.

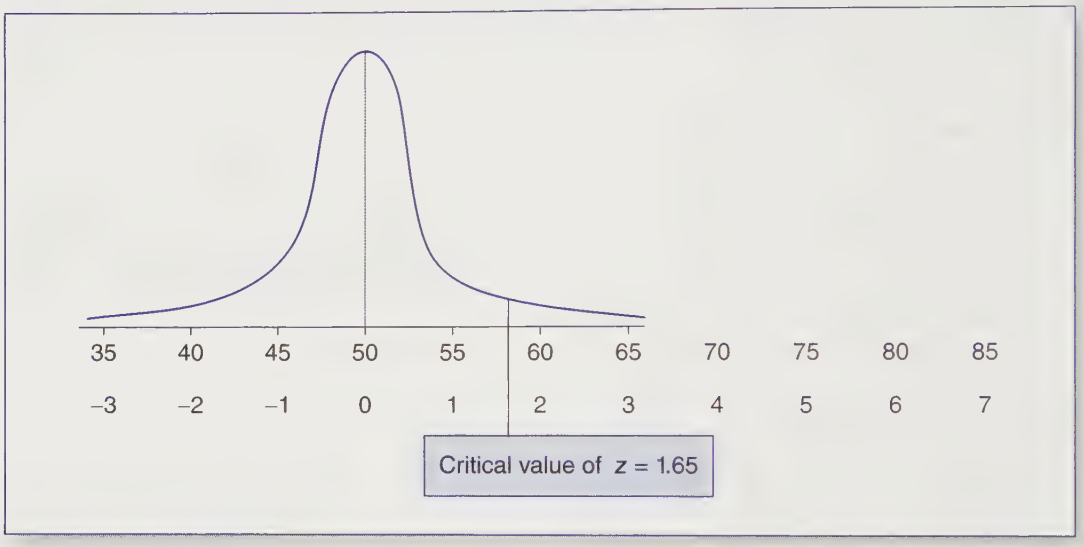
36. Why does sampling error decrease as the sample size increases?

    a. Sampling error is computed by dividing the *SS* for the population by *N*. If you increase *N*, you decrease sampling error.

    b. Sampling error is computed by dividing the population standard deviation by the square root of *N*. If you increase *N*, you decrease sampling error.

A major benefit of locating the center of the research hypothesis distribution of sample means, even if it is an *estimated* center, is that it allows researchers to quantify several other very important statistical concepts. This quantification process can be illustrated by "building a distribution." You have

already built one of the distributions that are necessary in Question 22. As you know from Question 22, the null hypothesis distribution of sample means is centered at a sample mean of 50 and a z score of 0. You also know that the spread of this null hypothesis distribution is determined by the standard error of the mean, $\frac{\sigma}{\sqrt{N}}$. You created this distribution in Question 22. This distribution is re-created for you below.

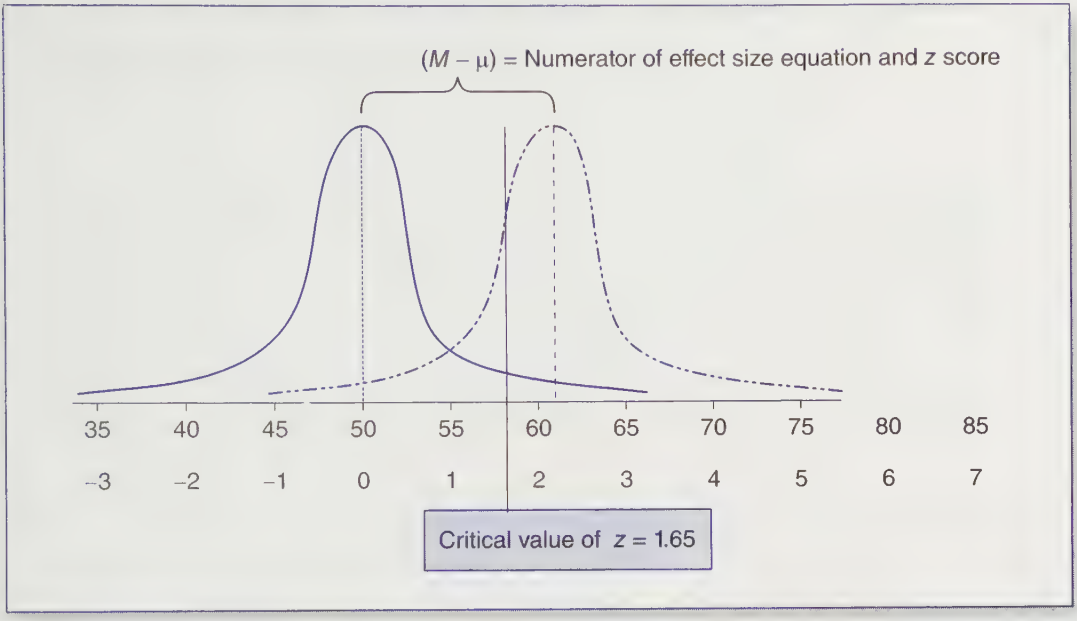


37. The second distribution you need to build is the distribution of the sample means *if the research hypothesis is true.* As mentioned earlier, the center of this distribution is determined by the actual mean of the sample. In this case, the actual sample mean was 61, meaning that the center of this distribution of sample means is at 61 on the raw score number line (and at the z score associated with a sample mean of 61, or 2.2). Draw a vertical line on the above figure at +2.2 to represent the center of the distribution of sample means if the research hypothesis is true. Remember that the only reason we can locate this curve's center is because we know the actual mean of the sample that received tutoring. We can't know this *before* we collect data.

38. The spread of this new distribution *is assumed* to be the same as that of the null distribution of sample means. This assumption simply means that we assume the variability around center after receiving tutoring is the same as it was without receiving tutoring (this is the homogeneity of variance assumption you evaluated in Question 1). Sketch in the distribution of sample means if the research hypothesis is true. Just try to duplicate the null curve's shape and spread but have it centered at 61 rather than at 50.

After you have completed sketching in the research hypothesis distribution of sample means, confirm that your figure looks like the following figure.

If your figure does not look like this one, determine what is wrong and fix it. Be sure you understand why the figure looks as it does. If you don't understand something, be sure to ask your instructor.



Previously, we said that "locating the center of the distribution of sample means *if the research hypothesis is true* allows researchers to quantify several other very important statistical concepts." Now, it's time to reap the benefits. To reap these benefits, you need to realize that both of the above curves are frequency histograms, and the areas under each respective curve represent all possible outcomes. The null distribution of sample means represents all possible outcomes (i.e., all possible sample mean values) if the null hypothesis is true. The research distribution of sample means represents all possible outcomes (i.e., all possible sample mean values) if the research hypothesis is true. By "cutting" these curves into different sections at the critical value of *z*, we can determine the probability of (a) rejecting a false null (i.e., statistical power), (b) not rejecting a false null (i.e., Type II error), (c) rejecting a true null (i.e., Type I error), and (d) not rejecting a true null. In the following figure, there is a distribution of sample means for the null hypothesis and a distribution of sample means for the research hypothesis. The two curves have been "separated" so that it is easier to see the distinct sections of each curve. The areas under each of these respective curves represent statistical power, Type II errors, Type I errors, and not rejecting a true null. Try to determine which areas under each curve below represent (a) rejecting a false null (i.e., statistical power), (b) not rejecting a false null (i.e., Type II error), (c) rejecting a true null (i.e., Type I error), and (d) not rejecting a true null.

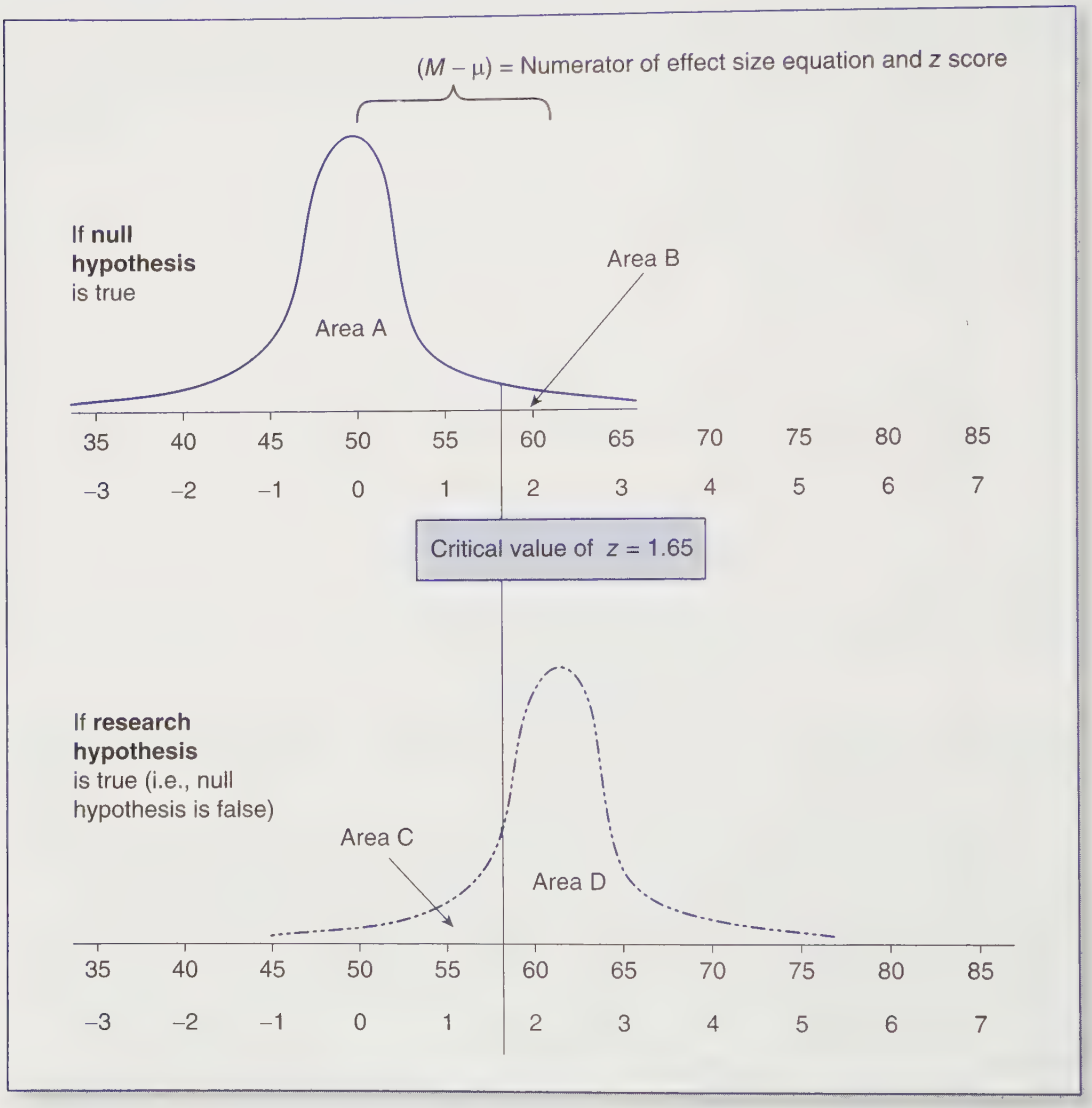
39. Match the area to the statistical outcome

Type I error (rejecting a true null): Area _____

Type II error (failing to reject a false null): Area _____

Statistical power (rejecting a false null): Area _____

Failing to reject a true null: Area _____



Effect Size

In the above figure, the mean difference between the peak of the "null" and "research" distributions of sample means represents the size of the effect that the IV (independent variable) had on the DV (dependent variable). In this case, it is the size of the effect that tutoring had on science test scores. As mentioned earlier, the size of the IV's effect is very important information when evaluating whether or not an IV is worth implementing. Researchers often want to quantify precisely how effective IVs are so that the relative effectiveness of different IVs can be compared. Comparing the relative effectiveness of IVs across different experiments is difficult for many reasons. For example, experiments often differ in their DVs (e.g., ACT or SAT scores) and/or their sample sizes. These differences make it

impossible to compare the "difference between the peaks" or $(M - \mu)$ across experiments. Instead, you must correct for these differences by computing *d*, a measurement of effect size that is not affected by differences in DVs or differences in sample size.

$$d = \frac{(M - \mu)}{\sigma}.$$

You should notice that the numerator is the "difference between the peaks." Dividing this difference by the standard deviation of the population in the study creates a common metric on which effect sizes can be compared across experiments with different DVs and/or different sample sizes.

40. In the preceding scenario, the sample mean was 61. Use this information and the fact that the population mean and standard deviation were 50 and 20, respectively, to compute the effect size of tutoring.

41. Use the information from your reading to interpret the effect size of tutoring using terms like *small, small to medium, medium, medium to large,* or *large.*

42. Interpret the effect size in terms of standard deviation units.

    The mean test score for the tutoring group was _____ standard deviations above the mean for the population that did not receive tutoring.

43. The process of significance testing yields only a decision as to whether the null hypothesis or the research hypothesis is more likely to be true. It does not indicate how effective the IV was in effecting the DV. In the above scenario, if the null hypothesis was rejected, you would only know that tutoring increased science test scores, but you would not know how effective the tutoring actually was. How would you determine how effective the tutoring was?

    a. Look at the *p* value; the lower the *p* value, the more effective the treatment.

    b. Look at the effect size; the greater the *d*, the more effective the treatment.

    c. Look at the standard error of the mean; the smaller the *SEM*, the more effective the treatment.

Activity 6.2: Critical Values, *p* Values, and the Null Hypothesis

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Describe what a *p* value represents
- Describe the relationship between critical values, *p* values, and the null hypothesis

1. One of the most sought-after skills by employers is the ability to work well with others. Unfortunately, local employers have told the local college that many of its graduates lack the necessary interpersonal skills for successful teamwork, and this deficit makes them difficult to

hire. To better prepare students for the workforce, the college hires Dr. U. B. Nice. Dr. Nice has successfully increased the interpersonal skills of many employees by implementing an 8-week program emphasizing active listening, communication, and emotion regulation. Although Dr. U. B. Nice has been successful with employees, he has never tried the program on college students. To determine if the program also improves the interpersonal skills of college students, Dr. Nice uses his program on 36 seniors who are randomly selected from the student body. After experiencing the 8-week program, each senior's interpersonal skills are measured with a standardized, structured interview. Interviewees are asked a variety of questions by the interviewer, and the interviewer scores each person on his or her responses. The resulting scores are interval/ratio. Other researchers have used this same structured interview with college students, and they have found that the distribution of interpersonal skills scores is normally distributed with a population mean of $\mu = 16$ and a standard deviation of $\sigma = 9$. After completing the program, the mean for this sample of 36 seniors was 19.20 ($SD = 8.24$).

This study meets all of the necessary assumptions. Match each of the statistical assumptions with the fact that suggests that assumption is met.

___c___ Independence

___d___ Appropriate measurement of the IV and the DV

___b___ Normality

___a___ Homogeneity of variance

a. The standard deviations of the sample and the population are similar.

b. The sample size used in the study is greater than 30.

c. The interviews in which the students' study skills were measured were conducted one student at a time under controlled conditions.

d. The IV in this study was a grouping variable that identifies how the sample is different from the population, and the DV is measured on an interval/ratio scale of measurement.

2. Write H_0 next to the one-tailed null hypothesis for this study and H_1 next to the one-tailed research hypothesis.

   a. ___H1___ The population of seniors who complete the program will have interpersonal skill scores higher than 16 ($\mu_{program} > 16$).

   b. ___H0___ The population of seniors who complete the program will not have interpersonal skill scores higher than 16 ($\mu_{program} \leq 16$).

   c. ___H1___ The population of seniors who complete the program will have interpersonal skill scores higher than 19.2 ($\mu_{program} > 19.2$).

   d. ___H0___ The population of seniors who complete the program will not have interpersonal skill scores higher than 19.2 ($\mu_{program} \leq 19.2$).

3. Dr. Nice took a sample of 36 seniors from a population with a mean of $\mu = 16$ and a standard deviation of $\sigma = 9$. The central limit theorem (CLT) allows you to describe the distribution of sample means if the null hypothesis is true. According to the CLT, what is the mean of the distribution of sample means if the null hypothesis is true?

   a. 16

   b. 30

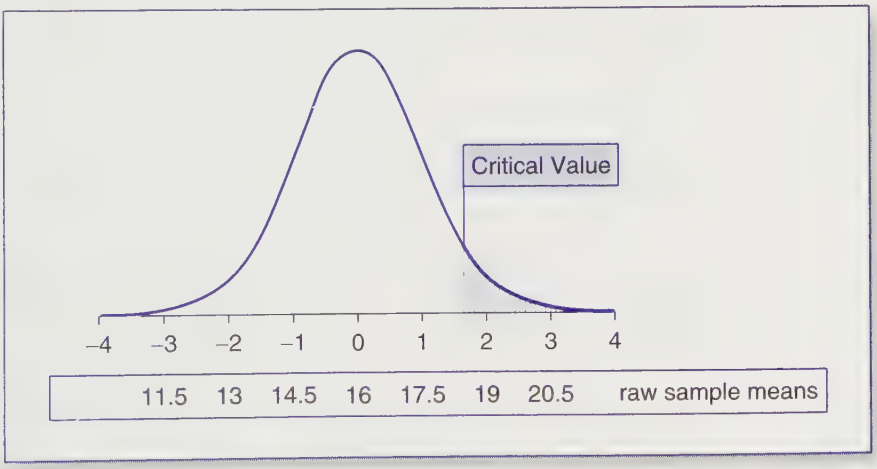
   c. 5

   d. 1.5

4. According to the CLT, what is the standard deviation of the distribution of sample means (i.e., SEM_p) if the null hypothesis is true?
   a. 16
   b. 30
   c. 5
   d. 1.5

5. According to the CLT, what is the shape of the distribution of sample means if the null hypothesis is true?
   a. Normally distributed
   b. The same as the original population of scores

6. The null distribution of sample means is drawn below. Make sure that the numbers on the distribution match your answers to Questions 3 through 5. The researchers want to use an alpha of .05. What is the critical value for this one-tailed test with an alpha of .05?
   a. 1.65
   b. 1.96
   c. 18.2



7. If the alpha value is .05. What is the probability that the researchers will make a Type I error if the null hypothesis is true?
   a. 0
   b. .01
   c. .05
   d. .95

8. Which of the *z* scores on the distribution could occur if the null hypothesis is true?
   a. Any *z* score could occur if the null hypothesis is true.
   b. Only the *z* scores in the critical region could occur if the null hypothesis is true.
   c. Only the *z* scores outside of the critical region could occur if the null hypothesis is true.

9. The z scores in the critical region are values that are
   a. likely to occur if the null hypothesis is true.
   b. unlikely to occur if the null hypothesis is true.

10. The critical value defines the critical region (the shaded area of the null distribution). What do you do if the computed z score (i.e., the obtained z score) is in the critical region?
   a. Reject the null hypothesis because that outcome could not occur if the null hypothesis is true
   b. Reject the null hypothesis because that outcome is unlikely to occur if the null hypothesis is true
   c. Fail to reject the null hypothesis because that outcome could not occur if the null hypothesis is true
   d. Fail to reject the null hypothesis because that outcome is unlikely to occur if the null hypothesis is true

11. Compute the obtained z score for this study.
   a. 1.5
   b. 1.65
   c. 2.13

12. Should Dr. Nice reject or fail to reject the null hypothesis?
   a. Reject the null hypothesis
   b. Fail to reject the null hypothesis

13. Which of the following is the best interpretation of the results of this study?
   a. The program definitely improved the interpersonal skills of seniors.
   b. It is likely that the program improved the interpersonal skills of seniors.

14. Dr. Nice obtained a z score of 2.13 for the sample of 36 seniors. What is the probability of obtaining a z score of 2.13 or higher if the null hypothesis is true? (Determine if you need the body or tail probability from the unit normal table to answer this question.)
   a. .9834
   b. .0166
   c. .7645
   d. .0343

15. Which of the following is a rule that indicates how you can use the p value (i.e., your answer to the previous question) to determine if you should reject the null hypothesis?
   a. Reject the null if $p > 1.65$
   b. Reject the null if $p < 1.65$
   c. Reject the null if $p < .05$
   d. Reject the null if $p > .05$

16. The null distribution of sample means represents all possible outcomes if the null hypothesis is true. In the null distribution, values that are farther from the population mean (or a z score of 0) are _____ likely than values that are closer to the population mean.
    a. less
    b. more

17. The p value is the probability of getting the obtained z score or one more extreme if the null hypothesis is true. (Hint: Remember that the distribution of sample means represents all possible outcomes if the null hypothesis is true.)
    a. True
    b. False

18. The p value is the probability of making a Type I error. (Hint: Look at Question 7.)
    a. True
    b. False

19. The p value is the probability that the null hypothesis is true.
    a. True
    b. False

20. You obtain a z score of 1.52. What is the p value?
    a. .9357
    b. .0643

21. You obtain a z score of 1.41 with a p value of .0793. Should you reject or not reject the null hypothesis?
    a. Reject
    b. Fail to reject

22. You obtain a z score that is not in the critical region; the p value must be
    a. greater than alpha (.05).
    b. less than alpha (.05).

23. You obtain a z score that is in the critical region; the p value must be
    a. greater than alpha (.05).
    b. less than alpha (.05).

24. You obtain a z score with a p value of .02; the z score must be
    a. in the critical region.
    b. outside of the critical region.

Activity 6.3: Statistical Power, Type I Error, and Type II Error

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Use two distributions of sample means to make probability statements about statistical power, Type I error, the probability of rejecting a false null, and Type II error
- Define statistical power, Type I error, the probability of rejecting a false null, and Type II error
- Describe how changing each of the following affects statistical power, Type I error, the probability of rejecting a false null, and Type II error:
  - Effect size of a treatment
  - Sample size of a study
  - Alpha level
  - Variability in a sample
- Apply the above relationships to novel scenarios and determine if researchers are being "dishonest" when presenting their statistical results
- Determine whether an alpha level of .05 or .01 is more appropriate by considering whether a Type I error or a Type II error is more "costly"
- Use the available statistics to determine which of several studies has produced the most promising results

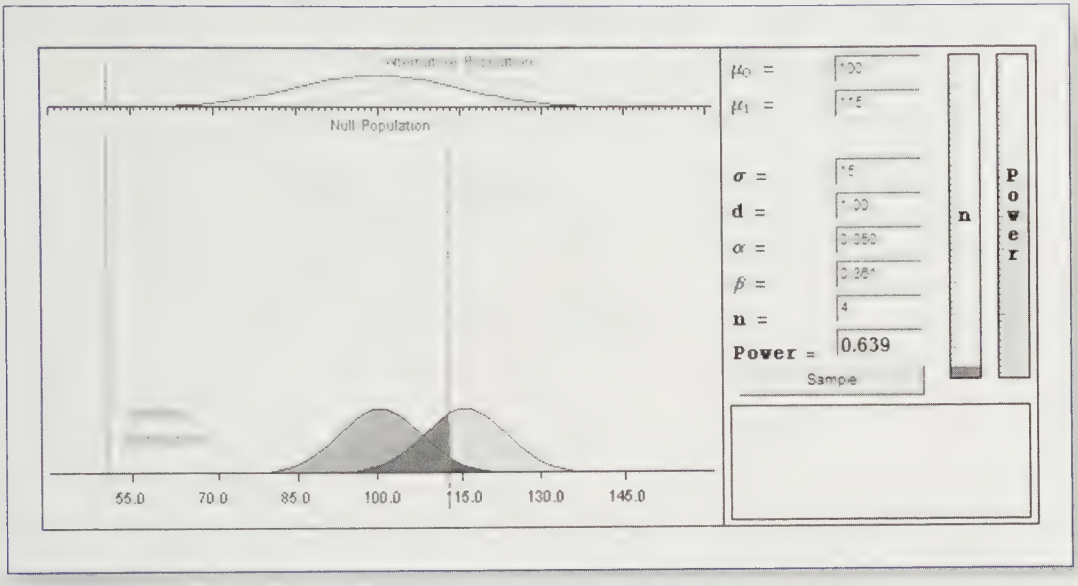
AN EXAMPLE: DETERMINING IF A NEW DRUG WORKS

A pharmaceutical company has developed a drug (New Drug) to help treat depression. Clinically depressed people on medication often have difficulty concentrating, and so the pharmaceutical company has developed this New Drug to increase depressed people's ability to concentrate. Currently, another drug (Old Drug) dominates the market. Suppose that it is known that people with depression who use the Old Drug have an average score of $\mu = 50$ on a mental concentration inventory, with a standard deviation of $\sigma = 10$. The mental concentration scores are measured on an interval/ratio scale and are normally distributed in the population. To test the effectiveness of the New Drug, a sample of 25 people with depression are given the New Drug, and their mental concentration inventory scores are recorded. Higher scores reflect greater mental concentration.

WISE Statistical Power Applet*

Go to the website http://wise1.cgu.edu and then click on "Statistical Power" under the WISE Applet heading on the left side of the page.

*©WISE Team. Used with permission.



In the Power Applet, type the following values into the boxes: $\mu_0 = 50$, $\mu_1 = 56$, $\sigma = 10$, $N = 25$, $\alpha = .05$. **Be sure to hit the ENTER key after typing each value.** When you are done, your graphs should look like those that follow.

1. The two population distributions at the top of the graphic represent the two populations of people who got the Old Drug (*blue line*) and the New Drug (*red line*). Researchers would not normally know the precise location of the New Drug population distribution, but it is provided here for teaching purposes. The two curves at the bottom represent the distribution of sample means if the null is true (*blue histogram*) and the distribution of sample means if the null is false (*pink histogram*). These are the two curves you created during Activity 6.1 on hypothesis testing. Notice how much narrower the *sample mean* distributions are relative to the *original population* distributions, which are displayed at the top of the graphic. Explain why this is not at all surprising.

Blue Curve

The shaded blue curve represents the sampling distribution assuming the null hypothesis is true (i.e., the New Drug did not increase mental concentration). The dark blue area shows the critical region for a one-tailed test. Because we selected an alpha of .05, the critical region is exactly 5% of the null hypothesis distribution.

2. What is the mean of this sampling distribution?
3. True or false: The mean of the null sampling distribution is the same as the mean of the Old Drug population.
4. What is the standard deviation of this sampling distribution (i.e., the standard error of the mean)?

   (Hint: $SEM = \frac{\sigma}{\sqrt{N}}$.)
5. True or false: The standard deviation of the null distribution of sample means is smaller than the standard deviation of the population.

6. What happens if a Type I error occurs in this study?
   a. The researcher concludes that the New Drug was more effective than the Old Drug, but it was not more effective.
   b. The researcher concludes that the New Drug was more effective than the Old Drug, and it was more effective.
   c. The researcher concludes that the New Drug was not more effective than the Old Drug, but it was more effective.
   d. The researcher concludes that the New Drug was not more effective than the Old Drug, and it was not more effective.
7. Which area of the graph represents the probability of making a Type I error?
   a. Light blue
   b. Dark blue
8. Which area of the graph represents the probability of correctly failing to reject the null hypothesis?
   a. Light blue
   b. Dark blue
9. For this study, if the researchers failed to reject the null hypothesis, what conclusion would they draw about the New Drug?
   a. It was not effective.
   b. It was effective.

Red Curve

The red curve represents the sampling distribution *assuming the null hypothesis is false* (the New Drug did increase mental concentration). For now, we've set the mean of this distribution as 56. As was explained in Activity 6.1, the precise location of this distribution is based on a single sample mean, and therefore the location should be considered an estimate. This distribution represents all possible sample means if the New Drug increases mental concentration to 56. You can see that the mean of this distribution is 56, and the standard deviation of this distribution is $\frac{10}{\sqrt{25}} = 2$ (the standard error of the mean).

10. What is the statistical power in the context of this study?
    a. The probability of correctly concluding that the New Drug *was* more effective than the Old Drug
    b. The probability of correctly concluding that the New Drug *was not* more effective than the Old Drug
    c. The probability of incorrectly concluding that the New Drug *was* more effective than the Old Drug
    d. The probability of incorrectly concluding that the New Drug *was not* more effective than the Old Drug
11. Which area of the graph represents statistical power?
    a. Everything to the right of the red line (*pink*)
    b. Everything to the left of the red line (*red*)

12. What happens if a Type II error occurs in this study?
    a. The researcher concludes that the New Drug was more effective than the Old Drug, but it was not more effective.
    b. The researcher concludes that the New Drug was more effective than the Old Drug, and it was more effective.
    c. The researcher concludes that the New Drug was not more effective than the Old Drug, but it was more effective.
    d. The researcher concludes that the New Drug was not more effective than the Old Drug, and it was not more effective.
13. Which area of the graph represents the probability of making a Type II error?
    a. Red
    b. Dark blue

The Curves and Numerical Values

In addition to showing the distributions graphically, the Applet also gives you exact numerical values for Type I error, Type II error, and statistical power. You can find these values listed on the right side of the display.

- The total area under each curve is equal to 1.
- Type I error is set by alpha (α). We set α at .05, and so we set Type I error at .05.
- Type II error is sometimes called beta error (β). For this study, it is .088.
- Statistical power is labeled "Power," and it is .912. Note that Power = $1 - \beta$.
- The effect size is *d*, and it is the difference between μ_0 and μ_1 divided by σ. (Note that this is similar to the *d* you calculated on previous activities.)

Hypothesis Testing

14. The hypothesis testing process is frequently used to determine whether a treatment is effective. You learned previously that hypothesis testing is based on evaluating a null hypothesis. The null hypothesis states that the treatment will not work at all, and therefore, the null hypothesis predicts that
    a. those who receive the treatment will have higher scores than will the population who did not receive the treatment.
    b. those who receive the treatment will have lower scores than will the population who did not receive the treatment.
    c. those who receive the treatment will have the same or lower scores compared to the population who did not receive the treatment.
15. If the mean of those who receive the treatment (i.e., the sample mean) is not exactly the same as the population mean, there are two possible explanations for the difference: (1) sampling error created the difference or (2) the treatment created the difference. If sampling error created the difference, the _____ would be true.

a. research hypothesis
b. null hypothesis

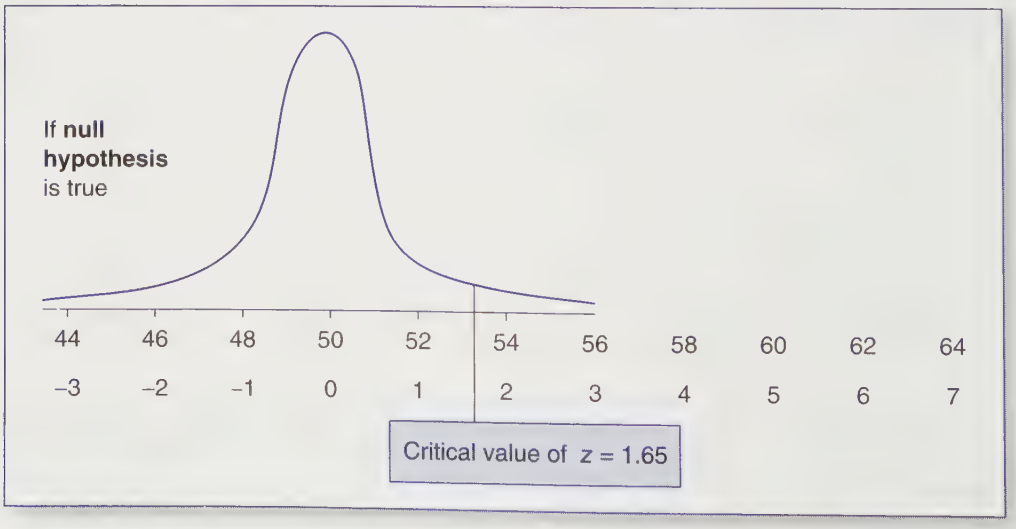
16. If the treatment created the difference, the _____ would be true.
    a. research hypothesis
    b. null hypothesis

17. The sole purpose of hypothesis testing is to determine whether the observed difference between the sample mean and the population mean was created by sampling error or by the treatment. How do researchers make this decision?
    a. If the z score for the sample mean is in the critical region, they conclude that the difference was created by sampling error.
    b. If the z score for the sample mean is in the critical region, they conclude that the difference was not created by sampling error and therefore it must have been created by the treatment.

In this problem, researchers are trying to determine if a new depression drug leads to higher mental concentration scores than another drug. Clinically depressed people who are using the Old Drug have a mental concentration score of 50 (i.e., $\mu = 50$), with a standard deviation of 10 (i.e., $\sigma = 10$). To test the New Drug, the researchers have to decide how much of the drug they should give to the people in the sample they selected ($N = 25$). Suppose they decided to give each person in the sample 80 mg of the New Drug each day. Further suppose that after taking the drug, the mean mental concentration score for those in the sample was 52 (i.e., $M = 52$).

18. Obviously, the sample mean of 52 is different from the population mean of 50. The question is whether this difference was likely to have been created by sampling error. Compute the z for the sample mean of 52 when $\mu = 50$, $\sigma = 10$, and $N = 25$.

19. Locate the z score you just computed on the following figure.



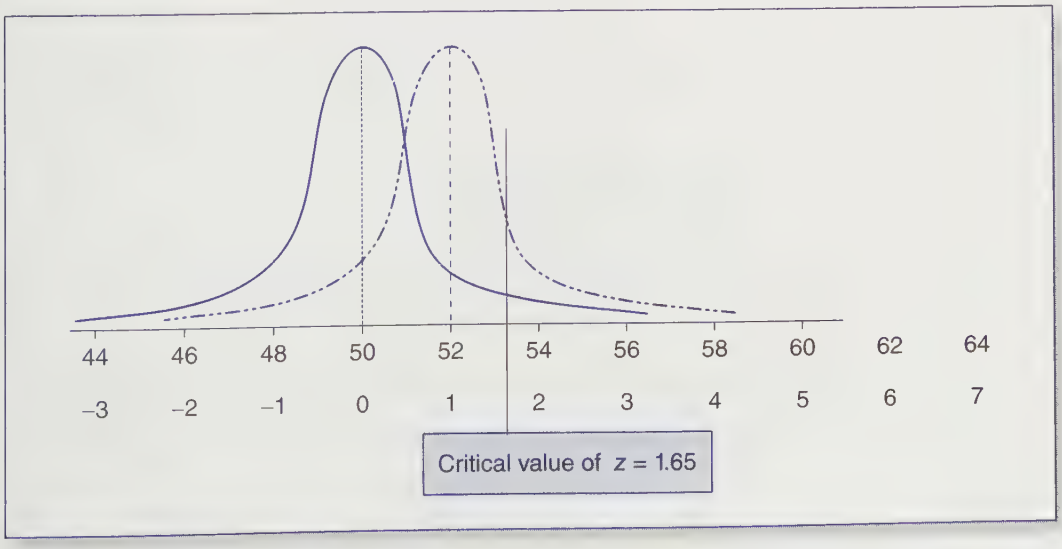
20. Would you reject or fail to reject the null hypothesis in this situation?
    a. Reject the null (i.e., the difference is probably not due to sampling error)
    b. Fail to reject the null (i.e., the difference is probably due to sampling error)

21. When $\mu = 50$, $\sigma = 10$, and $N = 25$, obtaining a sample mean of 52 (i.e., $M = 52$) leads to the conclusion that the treatment probably did not work because the difference between the sample mean and the population mean ($M - \mu = 2$) led to a *z* score of 1, which was small enough to have resulted from sampling error. However, it is also *possible* that the small difference was actually created by the treatment (i.e., the treatment worked, just not very well). If the treatment really did work but the researcher mistakenly concluded that it did not work, the researcher would have committed a
    a. Type I error.
    b. Type II error.
    c. Type III error.

After conducting a study in which the null hypothesis was not rejected, it is a good idea for researchers to recognize that they may have made a Type II error. Many researchers will want to estimate the probability that they made a Type II error. If the probability is too high, they may decide that it is necessary to conduct an additional study. While you will be using computers to perform the computations, it is important that you have a conceptual understanding of Type II errors.

To determine the probability that a Type II error was made (i.e., that researchers concluded the treatment did not work when it really did work), a second distribution of sample means is created. This second distribution of sample means is centered on the mean of the sample from the study; in this case, the sample mean was 52. If the treatment really improves mental concentration by 2 from 50 to 52 and we built a distribution of sample means *for those receiving the treatment,* it would be centered on 52. This second distribution is the research hypothesis distribution of sample means. The research hypothesis distribution of sample means has been added to the null distribution of sample means below.

22. Look at the following figure. Which curve is the null hypothesis curve, and which curve is the research hypothesis curve? Explain your reasoning.



23. The null hypothesis curve represents
    a. all possible sample means *if* the treatment does not work at all.
    b. all possible sample means *if* the treatment improves mental concentration by 2.
    c. all possible sample means.

24. The research hypothesis curve represents
    a. all possible sample means *if* the treatment does not work at all.
    b. all possible sample means *if* the treatment improves mental concentration by 2.
    c. all possible sample means.

25. If the treatment really improves mental concentration by 2 and you randomly selected a sample from the *research hypothesis distribution of sample means,* what sample mean would you expect for the sample? I would expect a sample mean of _____, which corresponds to a z score of _____.

26. If the sample you selected had the mean you expected above, would the z score corresponding to that mean lead you to reject the null hypothesis or fail to reject the null hypothesis?
    a. To reject the null because the z score would be in the critical region
    b. To fail to reject the null because the z score would not be in the critical region

At this point, you should notice that even if the treatment really does improve mental concentration by 2, most samples in the research hypothesis distribution of sample means would result in researchers failing to reject the null hypothesis, which would be a Type II error if the treatment really works.

27. This unfortunate fact is illustrated by the figure on p. 187 because most of the possible samples in the *research hypothesis curve* are located
    a. to the left of the critical value.
    b. to the right of the critical value.

28. The proportion of the *research hypothesis curve* that is to the left of the critical value is equal to the probability of committing a
    a. Type I error.
    b. Type II error.

29. The proportion of the *research hypothesis curve* that is to the right of the critical value is equal to the probability of
    a. *correctly* concluding that a treatment works, known as statistical power.
    b. *correctly* concluding that the treatment does not work.
    c. making a Type I error.

30. Enter the following values into the Power Applet: $\mu_1 = 52$, $\mu_0 = 50$, $\sigma = 10$, and $N = 25$. You will find that when the treatment effect is an improved mental concentration of 2, the probability of committing a Type II error is _____ and statistical power is _____.

31. When you add the probability of a Type II error and the value for statistical power together, you will always get exactly 1. Why is this true?
    a. They are the only two things that could occur if the research hypothesis is true, and so their probabilities must sum to 1.

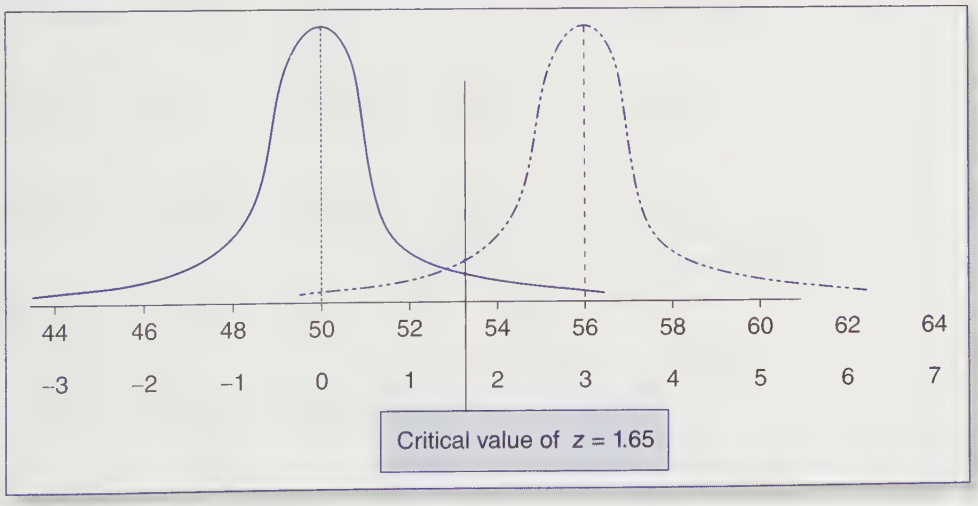
b. They are the only two things that could occur if the null hypothesis is true, and so their probabilities must sum to 1.

Factors That Influence the Hypothesis-Testing Process

Generally speaking, *researchers like their studies to have statistical power values of at least .80 and Type II error values less than .20.* The research study you just worked with had too little statistical power and a Type II error rate that was too high. As a consequence, the researchers decided to design a new study that might have more statistical power. Researchers can increase a study's statistical power by changing several things about the study. In this activity, you will learn how changing (a) the size of the treatment effect, (b) sample size, (c) alpha level, and (d) amount of variability created by measurement error each influences statistical power and Type II error rate.

Size of the Treatment Effect

One possible way researchers could increase the statistical power is to *increase the size of the treatment effect.* It is possible that if they gave the people in their sample 200 mg/day of the new antidepression drug rather than 80 mg/day, the treatment effect might be greater than the 2-point improvement. Suppose the higher dosage was given to a new sample, and further suppose that the sample's mean mental concentration was 56 rather than the population average of 50. The new research hypothesis distribution of sample means for this study is shown as follows:



32. What impact did increasing the treatment effect from 2 (52 – 50 = 2) to 6 (56 – 50 = 6) have on the proportion of the *research hypothesis curve* that was to the right of the critical value?

a. It increased the statistical power.

b. It decreased the statistical power.

33. In the computer Applet, change μ_1 from 52 to 56, and then record the exact values for statistical power and Type II error rate.

    The statistical power changed from .26 to _____.

    The Type II error rate changed from .74 to _____.

34. Summarize how increasing and decreasing the treatment effect changes each of the following by completing the following table:

If the treatment effect is	*Statistical power will (increase/decrease/ not change)*	*Type II error will (increase/decrease/ not change)*
Increased		
Decreased		

Increasing Sample Size

Of course, researchers cannot always increase the treatment effect. Suppose that the New Drug has too many negative side effects when taken at 200 mg/day. Another way researchers can increase the statistical power of a study is to increase the sample size of the study. Reset the values in the Power Applet to $\mu_0 = 50$, $\mu_1 = 52$, $\sigma = 10$, and $N = 25$. Record the values for statistical power and Type II error rate below.

35. The statistical power was _____, and the Type II error rate was _____. Be sure to look at the figure produced by the Applet. You will need to determine how the figure changes when N is increased.

36. Now change the sample size so $N = 100$, and compute the z score for a sample mean of 52 (i.e., the expected value if the improvement in mental concentration is 2). When the sample size was 25, the z score for a sample mean of 52 was 1. However, when the sample size was 100, the z score for a sample mean of 52 was _____.

37. The figure that reflects the change in sample size is on page 191. (Note that the values on the x-axis are different from those when $N = 25$; the graphs in the Applet are more accurate than those in this workbook.) You should notice that when the sample size is increased from 25 to 100, most of the sample means in the *research hypothesis curve* would lead researchers to

    a. reject the null hypothesis.

    b. fail to reject the null hypothesis.

38. Increasing the sample size will _____ the statistical power.

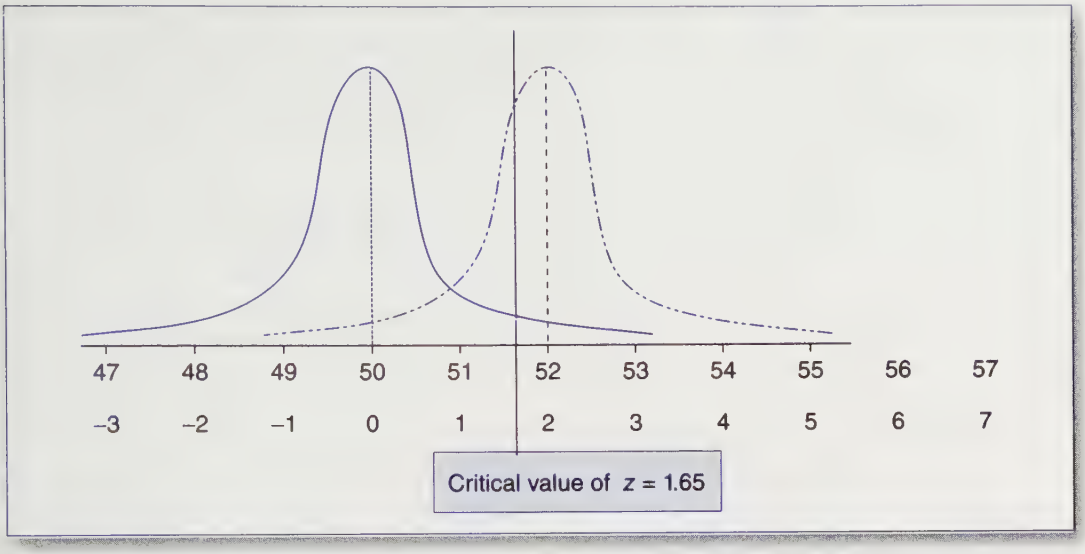
    a. increase

    b. decrease

39. Increasing the sample size will _____ the Type II error rate.

    a. increase

    b. decrease



40. As you _____ the sample size, you _____ sampling error and _____ statistical power.
    a. increase, decrease, decrease
    b. increase, increase, decrease
    c. increase, decrease, increase

41. Now enter the values for this study into the Power Applet, and record the exact values for statistical power and Type II error rate when the sample size is 100 rather than 25.

    When $N = 25$, statistical power is _____, and when $N = 100$, it is _____.

    When $N = 25$, Type II error rate is _____, and when $N = 100$, it is _____.

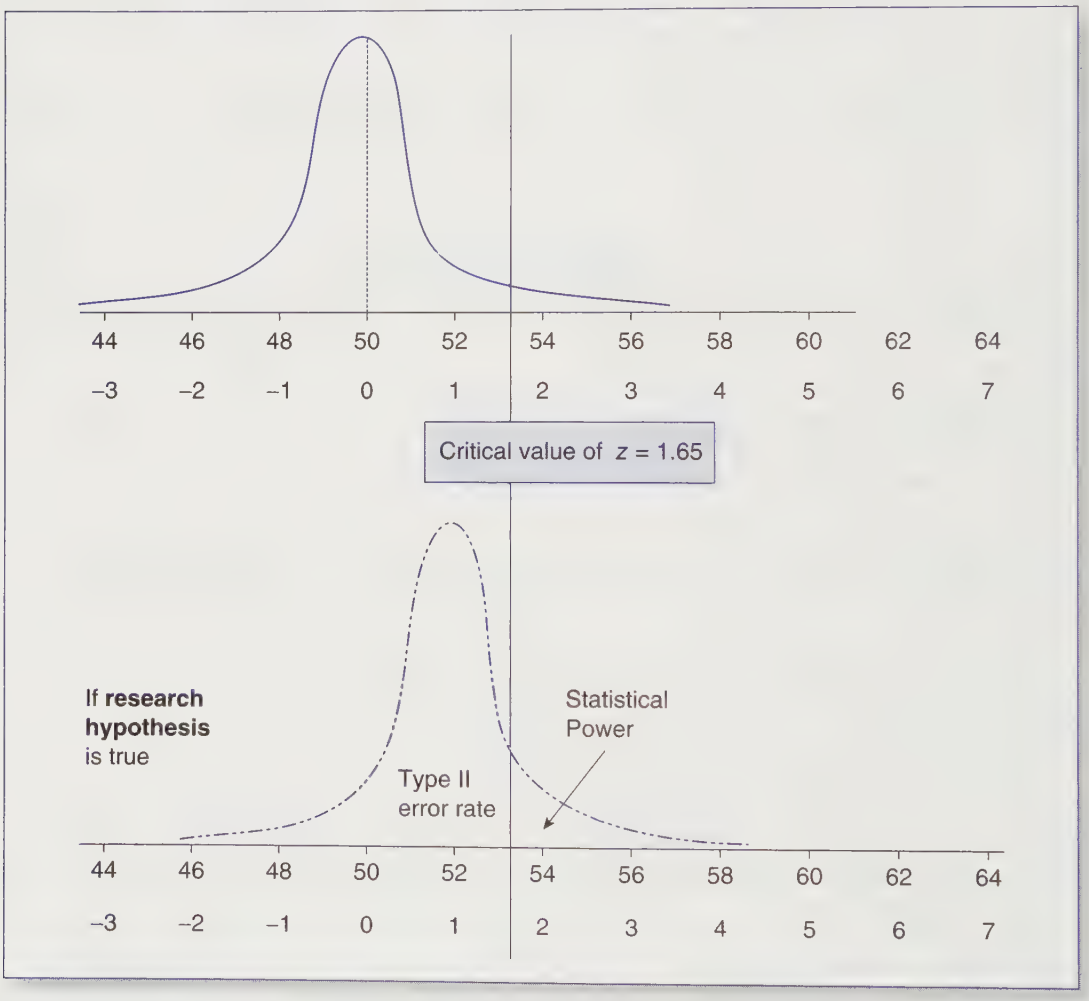
42. Summarize how increasing and decreasing the sample size affects each of the following by completing the following table:

If the sample size is	*Statistical power will (increase/decrease/ not change)*	*Type II error will (increase/decrease/ not change)*
Increased		
Decreased		

Changing the Alpha Level

Sometimes it is not possible to increase the sample size. Another way researchers can change the statistical power and Type II error rate of a study is by changing the alpha (α) level of the study. Essentially,

this means that the researcher changes the size of the critical region on the *null hypothesis curve*. If $\mu_0 = 50$, $\mu_i = 52$, $\sigma = 10$, $N = 25$, and $\alpha = .05$, the distribution of sample means for the null hypothesis curve would look like the top curve below. The area to the right of the critical value in this curve is 5% of the overall curve. The distribution of sample means for the research curve would look like the following bottom curve.



43. If researchers decreased the α level from .05 to .01 (i.e., moved the critical value line to the right), what impact would it have on the study's statistical power?

    a. It would increase the statistical power.

    b. It would decrease the statistical power; therefore, in this situation, the researchers would probably leave the α level at .05.

44. Researchers in psychology usually choose between α levels of .05 and .01. The α level the researchers choose has a direct impact on the Type I error rate. An α level of .05 has a 5% Type I error rate, and an α level of .01 has a 1% Type I error rate. What is a Type I error?

    a. A Type I error is saying the treatment works when it does work.

    b. A Type I error is saying the treatment works when it does not work.

45. Summarize how increasing and decreasing the α level affects each of the following by completing the following table:

If the α level is	*Statistical power will (increase/decrease/not change)*	*Type I error will (increase/decrease/not change)*	*Type II error will (increase/decrease/not change)*
Increased to .05			
Decreased to .01			

Reducing Variability by Reducing Measurement Error

Another way researchers could increase the statistical power of a study would be to improve the accuracy of their measurement. In this study, the dependent variable is the score on a mental concentration inventory. The scores are computed based on patients' answers to questionnaires. If some questions on the original questionnaire were confusing to the patients, it would produce measurement error. If the confusing questions were rewritten so that they are less confusing, it would reduce the measurement error, resulting in less variability in the study. In other words, the standard deviation of the population would be less, and the statistical power and Type II error rates would be affected.

46. Set the values for this research scenario at $\mu_0 = 50$, $\mu_1 = 52$, $\sigma = 10$, $N = 25$, and $\alpha = .05$. Record the values for statistical power and Type II error rate below:

    Statistical power is _____; Type II error rate is _____.

47. Now change the standard deviation of the population from $\sigma = 10$ to $\sigma = 5$. Record the values for statistical power and Type II error rate below:

    Statistical power is _____; Type II error rate is _____.

48. Summarize how decreasing the measurement error affects each of the following by completing the following table:

If the measurement error is	*Statistical power will (increase/decrease/not change)*	*Type II error will (increase/decrease/not change)*
Decreased		

Putting It All Together

49. The pharmaceutical company that developed the New Drug has invested many years and millions of dollars in developing this New Drug. Financially, it is very important that it brings this drug to market. Obviously, it is easier to convince doctors to prescribe this drug if the company can present scientific evidence that the New Drug is more effective than the Old Drug. Unfortunately, Phase I clinical trials suggest that the New Drug is only slightly more effective than the Old Drug. How could the sample size be manipulated to maximize the likelihood that their drug would be found to be "significantly better," even if the drug's actual effectiveness is only slightly better than that of the competitor's drug?
    a. Using a very large sample size can result in rejecting the null hypothesis even if the drug is only slightly better than the competitor's drug.
    b. Using a very large sample size can result in obtaining a very large effect size even if the drug is only slightly better than the competitor's drug.
    c. Using a very large sample size can result in failing to rejecting the null hypothesis even if the drug is much better than the competitor's drug.

50. Suppose you work for the company that manufactures the Old Drug and you find out that the above pharmaceutical company manipulated the sample size. You are still convinced that the Old Drug is a better option because it costs almost 75% less than the New Drug. How could you convince doctors that the difference between the effectiveness of the Old Drug and that of the New Drug is not really important, even though it was statistically significant in the studies conducted by the above pharmaceutical company?
    a. Compute the z scores and show them that the z is much larger for the New Drug than the Old Drug.
    b. Compute the exact p value and explain that this very small p value shows that the difference between the two drugs is actually very small.
    c. Compute the effect size and explain that the actual difference between the drugs is very small.

51. Suppose a pharmaceutical company has developed a new drug to treat a severe mental disorder that currently has no effective drug treatment available. The statistician responsible for analysis of the data suggests that they use an alpha level of .10 rather than .05 or .01. Is this a legitimate option in this case?
    a. Yes, the researchers are justifiably concerned about making a Type II error and using an alpha of .10 reduces the probability of making this type of error.
    b. Yes, the researchers are justifiably concerned about making a Type I error and using an alpha of .10 reduces the probability of making this type of error.
    c. No, the researchers are justifiably concerned about making a Type II error and using an alpha of .10 increases the probability of making this type of error.
    d. No, the researchers are justifiably concerned about making a Type I error and using an alpha of .10 increases the probability of making this type of error.

52. A researcher in another company reads the above study and sees that they used an alpha of .10 and decides to use an alpha of .10 in a study his company is currently conducting to test the effectiveness of a new cough drop. Is this a legitimate option in this case?
    a. In this case, the researchers are investigating a treatment for a relatively minor problem and treatments already exist. Using an alpha of .10 is a good idea because it will help maximize the statistical power and minimize the probability of making a Type II error.

b. Using an alpha of .10 is not acceptable because it would increase the Type I error rate. Although Type II errors would go down, there is little reason to be overly concerned with Type II errors in this study. Because current treatments exist for coughs, incorrectly failing to reject the null should not be the primary concern.

c. It does not matter if the researcher uses an alpha of .10 or .05. Either is acceptable as long as they use a very large sample size.

Suppose that the National Institutes of Mental Health have the money to fund two research programs. Although there is money to fund only two research programs, more than 30 research teams have applied for this grant money. A group of experts read all the grant applications and narrow it down to the five best ones. A number of factors are being considered, but one is the researchers' past success in developing treatments. Each of the researchers submitted information about the type of mental disorder and population they have studied, the sample size, the result of the significance test, and the effect size. Use the information provided in the following chart to answer the next question:

Study Number	*Type of Cancer/Population*	N	*Significance*	d
1	Depression in teenagers	500	$p < .05$ Reject H_0	0.25
2	Obsessive compulsive disorder in adults	15	$p < .05$ Reject H_0	0.92
3	Anxiety disorder in the elderly	210	$p < .05$ Reject H_0	1.2
4	Bipolar disorder in adults	430	$p < .05$ Reject H_0	0.57
5	Antisocial disorder in teenagers	9	$p > .05$ Fail to reject H_0	0.67

53. Based on these data, which two researchers would you recommend receive the research grant? Explain your answer using effect sizes, sample sizes, and the results of the hypothesis test.

Activity 6.4: Hypothesis Testing and Effect Size

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Conduct a hypothesis test using the *z* for a sample mean
- Distinguish between Type I and Type II errors

REDUCING TYPE I AND TYPE II ERRORS

In Activity 6.3, you learned about the effect of sample size, treatment effects, alpha levels, and variability on the probability of making a Type I and Type II error. Use that activity to answer the following two questions.

1. How can you decrease the probability of making a Type I error in a study?
   a. Increase the sample size (*N*)
   b. Decrease the sample size (*N*)
   c. Increase alpha (e.g., from .01 to .05)
   d. Decrease alpha (e.g., from .05 to .01)
   e. Make the treatment more powerful
   f. Make the treatment less powerful
   g. Increase the measurement error in the study
   h. Decrease the measurement error in the study
2. How can you decrease the probability of making a Type II error in a study? Select four. Use the letters from Question 1 to answer.

SCENARIO 1

College students frequently ask their professors how they should study for exams. Research on human memory and learning suggests that taking practice tests is a very effective learning strategy. In fact, several research studies suggest that taking practice tests is a more effective strategy than rereading chapters or reviewing class notes (Rowland, 2014). However, one of the obstacles to using the practice test strategy is that students don't always have access to practice tests. A college professor wondered if he could teach his students to test themselves while studying for exams. To answer this question, the college professor taught students to test themselves while they studied for his exams. For example, he taught students to test themselves by trying to summarize paragraphs they read and by explaining why each answer on previously completed classroom activities was correct. He taught these studying techniques to all 33 of his students (i.e., $N = 33$) and encouraged them to use these studying strategies to prepare for the exams in his course. The average score on the first exam in the course was $M = 76$, with a standard deviation of $SD = 11.7$. The professor wanted to know if his students' performance on this first exam was significantly better than the mean performance of his students from previous semesters on the same test. The mean score of his previous students on the test was $\mu = 74$, with a standard deviation of $\sigma = 10.9$.

3. This study meets all of the necessary assumptions. Match each assumption to the fact that suggests it is met.

   _____ Independence

   _____ Appropriate measurement of the IV and the DV

   _____ Normality

   _____ Homogeneity of variance

a. The standard deviation from the sample is similar to the standard deviation from the population.

b. The IV in this study is a grouping variable that identifies how the sample is distinct from the population, and the DV is measured on an interval/ratio scale of measurement.

c. The sample size in the study was greater than 30.

d. The students took the test under carefully monitored conditions that ensured students did not cheat.

4. Write H_0 next to the symbolic notation for the null hypothesis and H_1 next to the research hypothesis.

a. _____ $\mu_{\text{self-testing}} > 74$

b. _____ $\mu_{\text{self-testing}} < 74$

c. _____ $\mu_{\text{self-testing}} \geq 74$

d. _____ $\mu_{\text{self-testing}} \leq 74$

e. _____ $\mu_{\text{self-testing}} > 76$

f. _____ $\mu_{\text{self-testing}} < 76$

g. _____ $\mu_{\text{self-testing}} \geq 76$

h. _____ $\mu_{\text{self-testing}} \leq 76$

5. Write H_0 next to the verbal description of the null hypothesis and H_1 next to the research hypothesis.

a. _____The population of students who complete self-testing will have a mean test score greater than 74.

b. _____The population of students who complete self-testing will have a mean test score less than 74.

c. _____The population of students who complete self-testing will not have a mean test score greater than 74.

d. _____The population of students who complete self-testing will not have a mean test score less than 74.

6. Determine the critical value for this one-tailed *z* for a sample mean test (use $\alpha = .05$).

7. Compute the *z* for the sample mean test.

8. Based on the critical value of *z* and the obtained *z* score, the null hypothesis should

a. be rejected because the obtained *z* score is greater than the critical value.

b. be rejected because the obtained *z* score is less than the critical value.

c. not be rejected because the obtained *z* score is greater than the critical value.

d. not be rejected because the obtained *z* score is less than the critical value.

9. Which of the following is the best interpretation of these results?

a. The null hypothesis is definitely true.

b. There is insufficient evidence to reject the null.

10. Compute the effect size (*d*) for the professor's second study.

11. Determine if the effect size is small, small to medium, and so on.

12. Which of the following is a correct interpretation of the effect size?
    a. The sample mean is .18 standard deviations higher than the population mean.
    b. The sample mean is .18 standard deviations lower than the population mean.

13. Which of the following is the best summary of these results?
    a. The test-yourself study strategy did not lead to significantly higher exam performance than that produced by previous students, and the effect size was small.
    b. The test-yourself study strategy did lead to significantly higher exam performance than that produced by previous students, but the effect size was small.

14. You should have failed to reject the null hypothesis and concluded that the self-testing strategy did not improve performance. Whenever you *fail to reject a null hypothesis,* you should stop to consider the possibility that you *might* have committed a _____.
    a. Type I error
    b. Type II error

15. A Type II error occurs when a researcher concludes that the null hypothesis is _____ when the null hypothesis is _____.
    a. true; actually false
    b. true; actually true
    c. false; actually false
    d. false; actually true

16. Which of the following is the best description of a Type II error in this research scenario?
    a. Concluding that teaching students to self-test does increase exam scores when in fact it does not
    b. Concluding that teaching students to self-test does increase exam scores when in fact it does
    c. Concluding that teaching students to self-test does not increase exam scores when in fact it does not
    d. Concluding that teaching students to self-test does not increase exam scores when in fact it does

Because of the possibility of Type I and Type II errors, no single study definitively proves that a treatment has a reliable effect on scores. Therefore, careful researchers replicate studies to increase their confidence in the results. After an effect has been replicated, you can be more confident that your conclusion is accurate.

The professor was disappointed that the significance test indicated his self-testing strategy did not seem to help students perform better. However, because he was a careful researcher, he wanted to make sure he had not made a Type II error, so he planned to perform the study again. However, he wanted to strengthen the design of his next study by addressing as many of the potential problems that may have created a Type II error in his previous study as possible.

17. The professor could improve his future study by *decreasing the measurement error and/or increasing the procedural control to* decrease the variability in the study. Which of the following would accomplish this goal?
    a. He could grade the exams himself rather than having many different teaching assistants (TAs) grade the exams; doing so would reduce the "bad variability" in his study.
    b. He could continue to teach his students to use the self-test studying strategy in the hope that with more practice using the strategy, the students would get better at it.
    c. He could make his exams easier, so that his current students' scores would be higher than his previous students' scores.
    d. He could increase the number of students enrolled in his class.

18. The professor could improve his future study by *making the treatment more powerful*. Which of the following would accomplish this goal?
    a. He could grade the exams himself rather than having many different TAs grade the exams; doing so would reduce the "bad variability" in his study.
    b. He could continue to teach his students to use the self-test studying strategy in the hope that with more practice using the strategy the students would get better at it.
    c. He could make his exams easier, so that his current students' scores would be higher than his previous students' scores.
    d. He could increase the number of students enrolled in his class.

SCENARIO 2

The professor graded all of the future exams himself, and he continued to teach his students how to test themselves for the next 3 weeks. He also spent time encouraging them to use the self-test studying strategy outside class. After 3 weeks, he gave his current 33 students Exam 2, which was the same test he had given his previous students during previous semesters. The mean score of his current students on Exam 2 was $M = 81$, with a standard deviation of $SD = 11.3$. The mean score of his previous students on Exam 2 was $\mu = 77$, with a standard deviation of $\sigma = 10.4$.

19. This study meets all of the necessary assumptions. Match each assumption to the fact that suggests that assumption is met.

    _____ Independence

    _____ Appropriate measurement of the IV and the DV

    _____ Normality

    _____ Homogeneity of variance

    a. The students were tested in a procedurally controlled setting.
    b. The IV was a grouping variable, and the DV was measured on an interval/ratio scale of measurement.
    c. The standard deviations of the sample and the population were similar.
    d. The sample size was greater than 30.

20. Write H_0 next to the null hypothesis and H_1 next to the research hypothesis.
    a. _____The population of students who complete the practice test will have a mean test score greater than 77 ($\mu_{\text{practice test}} > 77$).
    b. _____The population of students who complete the practice test will have a mean test score less than 77 ($\mu_{\text{practice test}} < 77$).
    c. _____The population of students who complete the practice test will not have a mean test score greater than 77 ($\mu_{\text{practice test}} \leq 77$).
    d. _____The population of students who complete the practice test will not have a mean test score less than 77 ($\mu_{\text{practice test}} \geq 77$).
21. Determine the critical value for this one-tailed z for a sample mean test.
22. Compute the z for the sample mean test.
23. Based on the critical value of z and the obtained z score, the null hypothesis should
    a. be rejected.
    b. not be rejected.
24. Which of the following is the best interpretation of these results?
    a. The null hypothesis is definitely false.
    b. There is sufficient evidence to reject the null.
25. Compute the effect size (d) for the professor's second study.
26. Determine if the effect size is small, small to medium, and so on.
27. You should have found d to be .38, which is a medium effect size. Which of the following is the best interpretation of this effect size?
    a. Those students using the self-test studying strategy had scores that were .38 points better than those who did not use this strategy.
    b. Those students using the self-test studying strategy had scores that were .38 of a standard deviation higher than those who did not use this strategy.

28. The professor wanted to share his latest results with a colleague. Which of the following summaries (Option A or Option B) is the correct summary of his results? After you choose which one is correct, fill in the missing statistical values from his second study.

    Option A: The analysis of Exam 2 scores suggests that the self-test strategy is beneficial. The current students who were trained to test themselves while studying for exams had higher scores (M = _____) than the previous students who did not receive the training (μ = _____, σ = _____), z (N = _____) = _____, p = _____, d = _____.

    Option B: The analysis of Exam 2 scores suggests that the self-testing strategy is not beneficial. The current students who were trained to test themselves while studying for exams had scores that were *not* significantly higher (M = _____) than the exam scores of the previous students who did not receive the training (μ = _____, σ = _____), z (N = _____) = _____, p = _____, d = _____.

29. This time you should have rejected the null hypothesis and concluded the self-test strategy probably led to better performance than the previous students who did not use the strategy. Whenever you *reject a null hypothesis*, you should stop to consider the possibility that you *might* have committed a _____.

    a. Type I error

    b. Type II error

30. Which of the following best describes what a Type I error would be in this study?

    a. Saying that the self-testing strategy does not improve test scores when it does improve test scores

    b. Saying that the self-testing strategy does not improve test scores when it does not improve test scores

    c. Saying that the self-testing strategy does improve test scores when it does not improve test scores

    d. Saying that the self-testing strategy does improve test scores when it does improve test scores

31. If the null hypothesis is actually true, the probability of a Type I error is always equal to the

    a. critical value.

    b. obtained value.

    c. p value.

    d. alpha value.

32. When the null hypothesis is actually true and the α level is .05, researchers will make a _____ error _____ of the time.

    a. Type II; 5%

    b. Type I; 5%

    c. Type II; 1%

    d. Type I; 1%

33. A Type I error can be caused by sampling error. What can researchers do to reduce the potential problem of sampling error in their future studies? Select two.

    a. Use an alpha of .01 rather than .05

    b. Decrease the amount of variability in the sample by reducing measurement error

c. Increase the sample size

d. Use a different population that is more variable

SCENARIO 3

This third scenario involves a different college professor who was also interested in helping her students study and learn more effectively. She taught a large introductory psychology course with 312 students (i.e., $N = 312$). The professor wanted to try a new online introductory psychology textbook this semester to see if it would increase her students' scores relative to her previous students' scores on her comprehensive final exam. After the course was over, she compared the final exam scores. The mean final exam score of the current students who used the online text was $M = 74$, with a standard deviation of $SD = 13.5$. The mean final exam score of the previous students who used the traditional text was $\mu = 73$, with a standard deviation of $\sigma = 9.3$. The professor computed $z = 1.90$, $d = .11$, and rejected the null hypothesis.

34. Which of the following is the best interpretation of these results?

    a. The null hypothesis is definitely false.

    b. There is sufficient evidence to reject the null.

35. Which of the following is the best summary of these results?

    a. The students using the online text did not have significantly higher exam performance than those using the traditional text.

    b. The students using the online text did have significantly higher exam performance than those using the traditional text.

36. Which of the following is the best interpretation of the effect size?

    a. Those students using the online text had scores that were .11 points better than those using the traditional text.

    b. Those students using the online text had scores that were .11 of a standard deviation higher than those using the traditional text.

The results of the significance test indicate that the 1-point difference between the students using the online text and the students using the traditional text was unlikely to be due to sampling error (i.e., if the null hypothesis is actually true, there is only a 5% chance of the 1-point difference occurring due to sampling error). So, this decision is probably not a Type I error. The online text probably is better than the traditional text. However, the effect size you computed (i.e., $d = .11$) indicates that the online text is *only slightly better* than the traditional text. *Whenever you reject the null hypothesis and the effect size is small, you need to consider if the treatment is practically significant.* In other words, you need to consider if the treatment's effect is so small that it is not worth using. For example, suppose the online text costs $250 while the traditional text costs $50. While the significance test indicates that the online text is statistically better, the effect size indicates that it is only slightly better. Is having students perform .11 of a standard deviation better (i.e., in this case 1 point better) worth the additional $200 in cost? If your answer is "no," then you are saying that the treatment is not worth using. In terms of cost, the slightly better performance is not worth the substantially higher monetary cost. However, if the online text and the traditional text both cost $50, then you would probably decide that the online text is the better choice (i.e., the treatment is worth using). Depending on the practical costs, a small treatment effect can be very useful or not at all useful.

37. How it is possible that a treatment can actually be better than an alternative treatment but still not be worth using? How can you decide if a treatment is worth using or not?

    a. Look at the *p* value. If the *p* value is very low, then the results are very significant and you can be sure that the treatment worked and that it worked very well.

    b. Look at the *z* score. If it is in the critical region, you know that the results are statistically significant. If the treatment worked well enough to generate a *z* score in the critical region, you know that the treatment is worth using.

    c. Compute the effect size to determine if the treatment effect is large enough to warrant using the treatment.

Significance testing (also called hypothesis testing) and effect sizes have different purposes. The purpose of significance testing is to determine if a treatment works. Effect size determines the practical usefulness of a treatment. This distinction is important because many students ask, "Why do we need significance testing at all? Why don't we just use effect size?" Significance testing helps prevent us from adopting treatments that only appear to be better than other treatments due to sampling error. The following table may help you remember the different purposes of significance testing and effect size:

38. Why do researchers need both significance testing and effect size estimates? That is, what is the purpose of each statistical procedure? For example, suppose that you wanted to test the effects of a drug on people's memory scores. What different information would you get from the significance test and the effect size? Choose two.

    a. A significance test tells you if the difference between the sample mean and population mean is likely to be due to sampling error; an effect size tells you how much of a difference there was between the sample mean and the population mean.

    b. A significance tests tells you if the treatment worked; an effect size tells you how well the treatment worked.

    c. A significance test tells you if you should reject the null hypothesis; an effect size tells you if you should accept the research hypothesis.

39. How is it possible to fail to reject the null in one study when an effect size is large and to reject the null in another study when the effect size is small?

    a. If you have a very small sample size, you might fail to reject the null (due to a lack of statistical power) even though the effect size was large. In contrast, it is also possible to reject a null hypothesis when the effect size is very, very small if the sample size is very, very large.

    b. We compute effect sizes and hypothesis testing because they are completely independent of each other. A study with a large effect size is no more or less likely to be statistically significant than a study with a very small effect size.

Chapter 6 Practice Test

1. A high school principal knows that the average SAT on the writing portion of the test is $\mu = 436$ with a standard deviation of $\sigma = 110$. Troubled by these results, she institutes a new requirement for writing in all classes. Specifically, students are required to write at least one 8-page paper in every class. The following year, the average for a sample of 50 students in the senior class who took the SAT was 460.

   What is the null hypothesis for this study?

   a. $\mu_{writing} = 460$
   b. $\mu_{writing} = 436$
   c. $\mu_{writing} > 460$
   d. $\mu_{writing} < 460$
   e. $\mu_{writing} < 436$
   f. $\mu_{writing} > 436$

2. What is the research hypothesis for this study?

   a. $\mu_{writing} = 460$
   b. $\mu_{writing} = 436$
   c. $\mu_{writing} > 460$
   d. $\mu_{writing} < 460$
   e. $\mu_{writing} < 436$
   f. $\mu_{writing} > 436$

3. Which of the following situations would make the principal concerned that the assumption of independence was violated?

   a. All of the students were in the same school.
   b. Some of the students cheated on the SAT by looking at each other's papers.

4. Using the information in Question 1, draw a distribution of sample means assuming that the null hypothesis is true on a separate piece of paper. What is the mean and standard deviation of this distribution?

   a. 460; 15.56
   b. 436; 15.56
   c. 460; 110
   d. 436; 110

5. In the previous question, you generated a distribution of sample means assuming the null hypothesis is true. If you converted each of the sample means to a *z* score, what would the mean of this new distribution of sample means be after you converted the means to *z* scores?

   a. 0
   b. 1
   c. 436
   d. 460

6. What is the critical value for an alpha of .05 (one-tailed)?
   a. 1.96
   b. −1.96
   c. −1.65
   d. 1.65
7. Which of the following best describes the critical region?
   a. *z* values that are impossible if the null hypothesis is true
   b. *z* values that are impossible if the research hypothesis is true
   c. *z* values that are unlikely if the null hypothesis is true
   d. *z* values that are unlikely if the research hypothesis is true
8. Compute the *z* for a sample mean.
   a. 15.56
   b. 1.65
   c. 1.54
   d. .22
9. Should you reject or fail to reject the null hypothesis?
   a. Reject
   b. Fail to reject
10. Based on this study, did the writing program significantly improve test scores?
    a. The results suggest yes.
    b. The results suggest no.
11. Compute the effect size (*d*).
    a. 15.56
    b. 1.65
    c. 1.54
    d. .22
12. Which of the following best describes the size of this effect?
    a. Small
    b. Small to medium
    c. Medium
    d. Medium to large
    e. Large
13. Complete the following APA-style summary statement by filling in the missing values. You computed all of the missing values for previous questions. Scores on the writing portion of the SAT were not significantly higher after the writing program was implemented ($M =$ _____) than in the population before the program was implemented ($\mu =$ _____, $\sigma =$ _____), z ($N =$ _____) = _____, $p < .05$, $d =$ _____.

14. In the problem above, you decided if you should reject or fail to reject the null hypothesis by determining if the computed z was in the critical region. Another way is to determine if the p value is less than .05. What is the p value for this study?
    a. .0524
    b. .0618
    c. .4129
    d. .6245

15. Suppose that the principal repeated the study in 3 years after students had participated in the new writing program for the full 4 years of high school. What would this change likely do to the effect size?
    a. Increase it
    b. Decrease it

16. How can researchers increase the statistical power in their study? (Choose four.)
    a. Increase N
    b. Decrease N
    c. Increase the effect size
    d. Decrease the effect size
    e. Increase the population variability
    f. Decrease the population variability
    g. Use a larger alpha (e.g., .05 rather than .01)
    h. Use a smaller alpha (e.g., .01 rather than .05)

17. How can researchers decrease the Type I error rate in their study?
    a. Increase N
    b. Decrease N
    c. Increase the effect size
    d. Decrease the effect size
    e. Increase the population variability
    f. Decrease the population variability
    g. Use a larger alpha (e.g., .05 rather than .01)
    h. Use a smaller alpha (e.g., .01 rather than .05)

18. When you reject the null, what type of error might have you made?
    a. Type I
    b. Type II

19. When you fail to reject the null, what type of error might have you made?
    a. Type I
    b. Type II

REFERENCES

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463.

CHAPTER 7

Single-Sample *t* Test

LEARNING OBJECTIVES

After reading this chapter, you should be able to do the following:

- Explain when a single-sample *t* should be used rather than a *z* for a sample mean
- Explain why the *z* for a sample mean is superior to the single-sample *t* test but rarely used
- Write one- and two-tailed null and research hypotheses using population parameters and words
- Compute the degrees of freedom and define the critical region for one- and two-tailed single-sample *t* tests
- Explain why a *z* for a sample mean test and a single-sample *t* test have different critical regions
- Compute a single-sample *t* by hand and using SPSS
- Determine whether or not you should reject the null hypothesis
- Compute and interpret an effect size (*d*)
- Interpret the SPSS output for a single-sample *t*
- Summarize the results of the analysis using American Psychological Association (APA) style

SINGLE-SAMPLE *t* TEST

In the last chapter, you learned to use the *z* for a sample mean statistic to test a null hypothesis. To compute the *z* for a sample mean, you must know the *population's* standard deviation (σ). However, in most research situations, the population's standard deviation is not known. In this chapter, you will learn how to use the single-sample *t* statistic when you don't know the population standard deviation.

Like the *z* for a sample mean, the single-sample *t* statistic is used to determine if the difference between a sample mean and a population mean is likely to be due to sampling error. The *z* for a sample mean and the single-sample *t* are different because the single-sample *t* test uses the standard deviation of the sample (*SD*) rather than the standard deviation of the population to compute an *estimate* of expected sampling error. If you know the population standard deviation (σ), you should always do a *z* for a sample mean, but when you don't, you have to do a single-sample *t* test. *Almost* everything else is the same between the single-sample *t* and the *z* for a sample mean.

Reading Question

1. One of the differences between a *z* for a sample mean test and a single-sample *t* test is that
   a. the *z* for a sample mean is used when you don't know the population standard deviation.
   b. the single-sample *t* test uses the sample standard deviation to compute an estimate of the typical amount of sampling error.

You can also use the single-sample *t* statistic when you need to determine if a sample mean is significantly different from any number of "theoretical" interest. For example, you could compare the average amount of weight that a group of people lost while on a weight loss program to 0, which would represent no weight loss at all. In the weight loss example, the comparison value of 0 has a theoretical rationale because if a group of dieters' average weight loss does not exceed 0 pounds, it would suggest that the program does not work. Hypothesis testing with the single-sample *t* test can determine whether the weight loss program produced significantly more weight loss than 0 pounds. In this case, 0 pounds would be the value expected if the weight loss program did not work at all. It would be the value expected if the null hypothesis were true.

Reading Question

2. A researcher wants to assess people's knowledge of a topic by using a 10-question true/false test. Which value could be of theoretical interest to this researcher and, therefore, function as a null hypothesis test value?
   a. The researcher might compare the mean number of correct answers to 10, the number that represents perfect performance on the test.
   b. The researcher might compare the mean number of correct answers to 5, the average number of correct answers people would get if they were simply guessing on every question of the test.
   c. The researcher could use either "10" or "5" in this situation because both values represent a level of performance that might be of theoretical interest to the researcher.

CONCEPTUAL INFORMATION

The single-sample *t* test is logically identical to the *z* for a sample mean. Both tests compute the deviation between a sample mean and some expected value, which is either a population mean or a value of theoretical interest to the researcher. This deviation is the numerator of both the *z* test and the *t* test. The denominator of both tests is a value representing "typical sampling error." The minor computational difference between the *z* test and *t* test is in the computation of typical sampling error. Table 7.1 highlights the similarities and differences between these two significance tests.

When using the *z* test, the standard deviation of the population is used to compute the expected amount of sampling error (i.e., the denominator of the test, which is often called the **error term**). Specifically, sampling error is computed using the formula for the standard error of the mean $\left(SEM_p = \frac{\sigma}{\sqrt{N}}\right)$. In some situations, however, researchers may not know σ, the population's standard

deviation. In these situations, researchers compute the standard deviation of the sample data and use it to compute an estimate of expected sampling error. This procedure is one of the ways the single-sample *t* test differs from the *z* test. Specifically, sampling error is computed using the formula for the estimated standard error of the mean $\left(SEM_s = \frac{SD}{\sqrt{N}}\right)$ when using a sample *t* test.

Table 7.1 Similarities and Differences Between the *z* Test and *t* Test

	z for a Sample Mean Test	*Single-Sample* t *Test*
Formula	$z = \frac{M - \mu}{SEM_p}$	$t = \frac{M - \mu}{SEM_s}$
Computation of sampling error (standard error of the mean)	$SEM_p = \frac{\sigma}{\sqrt{N}}$	$SEM_s = \frac{SD}{\sqrt{N}}$

Reading Question

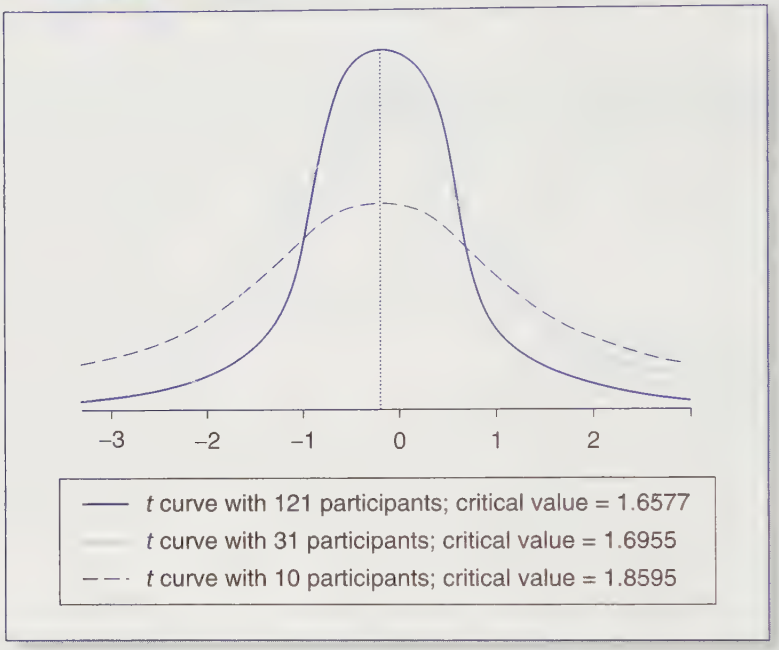
3. The only computational difference between the *z* for a sample mean formula and the single-sample *t* formula is the way
   a. the numerator is computed.
   b. typical sampling error is computed.

You should recall that critical regions define which values (*z* or *t* values) result in rejecting the null hypothesis. Another difference between *z* and *t* tests is the critical values used for each test. While the critical values for *z* tests are always the same, the critical values for *t* tests change based on sample size. For example, the critical value for a one-tailed*z* test with $\alpha = .05$ is always +1.65 *or* −1.65. Stated differently, the *z* scores of +1.65 *or* −1.65 will always be the points that "cut" the 5% of the *z* scores in the critical region from the 95% of *z* scores in the body of the *z* distribution. Conversely, the critical values for a one-tailed *t* test with $\alpha = .05$ will change based on how many people are in the sample. The reason for this is that the distribution of *t* values is not always normal in shape. When the sample size is small, the curve of *t* values is not shaped like a normal curve. In fact, *the shape of the* t *curve is different for every sample size*. Figure 7.1 illustrates how every sample size creates a different *t* curve by displaying the different *t* curves for three different sample sizes. The important consequence of the *t* curves having a different shape for every sample size is that the critical values that define the critical region are different for every sample size (i.e., *N*). Therefore, studies with different sample sizes will have different critical values even if both use one-tailed tests with $\alpha = .05$. As *N* increases, the *t* critical value is closer to zero.

Reading Question

4. The critical value for a one-tailed *z* test with $\alpha = .05$ is always
   a. +1.65 or −1.65.
   b. 0.
   c. +1.96 or −1.96.

Figure 7.1 The Shape of the *t* Distribution for Three Different Sample Sizes



Reading Question

5. The critical value for a one-tailed *t* test with $\alpha = .05$
   a. is the same as the critical value for a similar *z* test.
   b. changes based on the size of the sample (i.e., *N*) being used.

Reading Question

6. The critical value for a *t* test will get _____ as sample size increases.
   a. farther from zero
   b. closer to zero

Reading Question

7. As the sample size increases, the critical region for a *t* test
   a. gets larger (more than 5% of the distribution).
   b. gets smaller (less than 5% of the distribution).
   c. stays the same size (equal to 5% of the distribution), but its location changes.

The reason why the *t* curve has a different shape for every sample size is related to the fact that random sampling error changes so readily with sample size. Intuitively, it probably makes sense to you that larger random samples have less sampling error than do smaller random samples. After all, larger samples contain more of the population in them, and that implies less of a difference between the sample and the population.

Reading Question

8. When performing a *t* test, increasing the sample size will _____ the amount of sampling error expected. This means the denominator of the *t* test will be _____.

   a. increase; smaller
   b. increase; larger
   c. decrease; smaller
   d. decrease; larger

Aside from the differences in the computation and the critical value, the steps in hypothesis testing are the same for a single-sample *t* as with the *z* for a sample mean. The following example illustrates these steps.

ONE-TAILED SINGLE-SAMPLE *t* TEST EXAMPLE

College students in the United States are taking on student loan debt at record levels. As of 2013, the Consumer Financial Protection Bureau estimated that the total amount of federal student loan debt in the United States was over $1.2 trillion. Although there are many reasons for the increasing student loan debt, research on a common decision-making heuristic known as the optimism bias (Weinstein, 1980) suggests that one factor contributing to increasing student loan debt might be that students are overly optimistic about what their salaries will be after graduation. A financial aid counselor wants to see if students are overly optimistic about their future salaries. She knows that the distribution of starting salaries for psychology majors graduating from her school is normally distributed with a mean of $36,000 or $\mu = \$3{,}000$ a month. To determine if psychology students are overly optimistic about their potential salaries, the counselor obtains a random sample of 15 psychology majors from her school and asks each one of them individually what they expect their monthly salary will be in their first job after graduation. The 15 students' expected starting monthly salaries were 3,500, 4,200, 2,500, 3,200, 2,700, 2,000, 2,500, 4,100, 5,000, 4,200, 2,900, 3,900, 4,000, 5,000, and 5,500, which is a sample mean of $M = 3{,}680$. Previous research on the optimism bias (e.g., Weinstein, 1980) suggests that students will be overly optimistic in their estimates, and so the counselor chooses to do a one-tailed test. She also chooses an alpha of .05. Is the students' mean estimated starting salary significantly higher than the actual starting salary for psychology majors?

Step 1: Examine the Statistical Assumptions

The first assumption of *data independence* is met because the counselor collected the data carefully, not allowing any participants to influence other participants' answers. The second assumption of *appropriate measurement of the IV and the DV* is met because the DV, expected monthly salary, is measured on an interval/ratio scale and quantifies how much more (or less) each participant expects to make, and the IV, psychology students' expectations versus actual salaries, identifies a group of students' salary expectations versus a population's actual salaries. The third assumption that *the distribution of sample means will have a normal shape* is also met because the population of psychology majors' expected salaries is normally shaped.

When performing a single-sample *t* test, the last assumption of *homogeneity of variance* is more difficult to assess. The assumption is that the sample standard deviation (*SD*) is similar to the population standard deviation (σ), but when you are doing a single-sample *t* test, *you don't know the population standard deviation* (σ). So, researchers typically assess this assumption by using their knowledge of the variables they are studying. Is the "treatment" being applied to the sample likely to change the *variability* of scores? In this study, the counselor is only asking students how much money they expect to

make. This question, if asked carefully, is unlikely to make students' answers more or less variable, so it is safe to assume that the sample standard deviation will be similar to the population standard deviation. However, if the counselor were to conduct the study poorly, perhaps by asking a leading question like "Many students overestimate what their starting salaries will be. How much do you expect to make?" Asking the question this way could change students' answers, making every student lower his or her estimate, which could cause the sample's salary estimates to be less variable than the population's salary estimates, which would violate the homogeneity of variance assumption. Again, this example illustrates that poor research methodology can really mess up your statistics. In this study, the counselor worded the question carefully so the homogeneity of variance assumption was probably met.

Reading Question

9. You use a single-sample *t* when
   a. the IV defines two independent samples and the DV is measured on an interval/ratio scale.
   b. the IV defines two matched samples and the DV is measured on an interval/ratio scale.
   c. the IV defines one sample, the DV is measured on an interval/ratio scale, and the DV is measured twice on that same sample.
   d. the IV defines one sample and the DV is measured on an interval/ratio scale and you do not know the population standard deviation.
   e. the IV defines one sample and the DV is measured on an interval/ratio scale and you do know the population standard deviation.

Reading Question

10. When performing a single-sample *t* test, which of the following assumptions is the most difficult to assess?
    a. Data independence
    b. Appropriate measurement of the IV and the DV
    c. Normality of the distribution of sample means
    d. Homogeneity of variance

Step 2: State the Null and Research Hypotheses Symbolically and Verbally

In a one-tailed significance test, the researchers make a specific prediction about whether the sample mean will be higher or lower than the population mean, and therefore, there is one critical region. In this case, the *research hypothesis* would be that *the population of psychology majors' mean estimated starting salary is higher than $3,000/month* and therefore, the critical region would be on the positive side of the distribution. The *null hypothesis* would be that *the population of psychology majors' mean estimated starting salary is not higher than $3,000/month*. The null and research hypotheses will always be mutually exclusive and will cover all possible outcomes.

Reading Question

11. One-tailed significance tests
    a. have one critical region that is either on the positive or negative side of a distribution.
    b. have two critical regions, one on the positive side and one on the negative side of a distribution.

Reading Question

12. When using a one-tailed significance test, if the research hypothesis predicts an increase (or positive change), the critical region will be on the _____ side of the distribution.

a. negative

b. positive

You should notice two things about the null and research hypotheses for a one-tailed significance test. First, because they are mutually exclusive, one of them has to be correct. Second, the one-tailed hypotheses indicate that researchers will only reject the null hypothesis if the *t* value is far from zero and if it is in the direction stated in the research hypothesis. In this case, the researchers would only reject the null hypothesis if the *t* value is far from zero on the positive side of the *t* distribution. A *t* value in a positive critical region would indicate that the sample mean is significantly higher than the actual value of $3,000/month. In this example, a *t* value that is far from zero in the negative direction would result in failing to reject the null hypothesis. The symbolic and verbal representations of the research and null hypotheses are presented in Table 7.2.

Table 7.2 Symbolic and Verbal Representations for One-Tailed Research and Null Hypotheses for Single-Sample *t* Test

	Symbolic	*Verbal*	*Mean Difference Created by*
Research hypothesis (H_1)	H_1: $\mu_{\text{psych majors}} > 3{,}000$	The mean estimated starting salary of the population of psychology majors *is higher* than the actual mean starting salary ($3,000/month).	Psychology majors' overestimating their starting salaries
Null hypothesis (H_0)	H_0: $\mu_{\text{psych majors}} \leq 3{,}000$	The mean estimated starting salary of the population of psychology majors *is not higher* than the actual mean starting salary ($3,000/month).	Sampling error

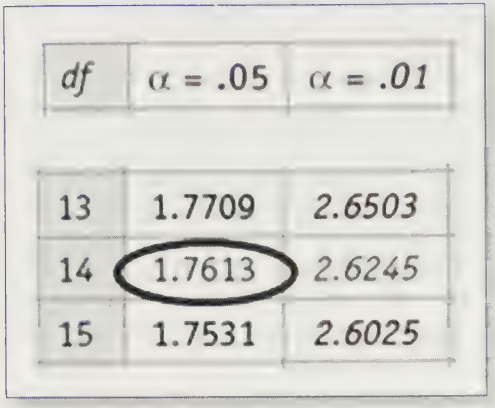
Step 3: Use Sample Size to Compute Degrees of Freedom and Define the Critical Regions

In the previous discussion about critical regions, you learned that as sample size increases, the critical value gets closer to zero. To locate the exact critical region, you will need the study's sample size as well as a table of critical *t* values (Appendix B). Once you know a study's sample size, you need to compute its **degrees of freedom (*df*)**. For the single-sample *t* test, the *df* formula is $df = N - 1$. So, in this case,

$$df = N - 1 = 15 - 1 = 14.$$

Then, you use the critical *t* values table in Appendix B to find the specific *t* critical value when $df = 14$. In this case, you should find the *df* of 14 in the left-most column of the one-tailed table and then find the critical value in that row under the .05 heading. Figure 7.2 illustrates how to use Appendix B.

Figure 7.2 Finding the Critical *t* Value



df	$\alpha = .05$	$\alpha = .01$
13	1.7709	2.6503
14	1.7613	2.6245
15	1.7531	2.6025

In this case, the critical *t* value is 1.7613. The research hypothesis predicted that our sample mean would be greater than our population mean (i.e., a positive difference for $M - \mu$). Thus, the critical region is on the positive side of the distribution, and the critical region starts at +1.7613. Therefore, the null hypothesis should be rejected if the obtained *t* we calculate in the next step is equal to or greater than +1.7613, expressed mathematically as *t obtained* ≥ +1.7613.

Reading Question

13. The degrees of freedom (*df*) for a single-sample *t* test are computed as
    a. *N*.
    b. $N - 1$.

Reading Question

14. The degrees of freedom are used to
    a. compute the single-sample *t*.
    b. determine the critical value.

Reading Question

15. If the research hypothesis indicates that scores will decrease, the critical value will be
    a. positive.
    b. negative.

Step 4: Compute the Test Statistic (Single-Sample *t* Test)

4a. Compute the Deviation Between the Sample Mean and the Population Mean

Begin by computing the deviation between the sample mean and the population mean. The mean estimated monthly salary from the 15 psychology majors was $M = 3{,}680$, and the actual monthly salary was $\mu = 3{,}000$. So, the mean difference is 680.

$$(M - \mu) = (3{,}680 - 3{,}000) = 680.$$

4b. Compute the Sampling Error Expected

To interpret the observed deviation of 680, you need to determine how much of a difference you would expect between *M* and μ due to sampling error. Sampling error is estimated by computing the estimated standard error of the mean $\left(SEM_{\bar{X}} = \frac{SD}{\sqrt{N}}\right)$. To compute sampling error, you need to know the standard deviation of the sample (*SD*) and the size of the sample (*N*). In this problem, the standard deviation of the sample is not given to you. Therefore, you must compute it from the sample data. As you may recall from Chapter 3, computing the *SS* (i.e., the sum of squared deviations) is the first step

to computing *SD*. The computational formula for *SS* is $SS = \sum X^2 - \frac{(\sum X)^2}{N}$. You need to sum all of the scores (*X*s) to find $\sum X = 55{,}200$. Then, you need to take every score (*X*) and square it, then sum all of the squared scores (X^2) to find $\sum X^2 = 218{,}240{,}000$. There were 15 scores, so $N = 15$. Make sure you can find each of these values and then find *SS*. The *SS* is computed as shown below.

$$SS = \sum X^2 - \frac{(\sum X)^2}{N} = 218{,}240{,}000 - \frac{(55{,}200)^2}{15} = = 15{,}104{,}000.$$

The standard deviation is computed by dividing the *SS* by $N - 1$ and taking the square root of the quotient.

$$SD = \sqrt{\frac{SS}{N-1}} = \sqrt{\frac{15{,}104{,}000}{14}} = 1{,}038.680.$$

And finally, the standard error of the mean is computed by dividing the standard deviation by the square root of the sample size.

$$SEM_{\bar{x}} = \frac{SD}{\sqrt{N}} = \frac{1{,}038.680}{\sqrt{15}} = 268.186.$$

In this case, the typical amount of sampling error was computed as 268.186. Therefore, sample means will typically be 268.186 away from the population mean due to sampling error.

Reading Question

16. For this example, the observed difference between the sample mean and the population mean was _____; the difference expected due to sampling error was _____.

a. 268.186; 680

b. 680; 268.186

4c. Compute the Test Statistic (Single-Sample *t* Test)

Dividing the observed difference between the sample mean and the population mean by the estimate of average sampling error yields the obtained *t* value.

$$t = \frac{M - \mu}{SEM_{\bar{x}}} = \frac{3{,}680 - 3{,}000}{268.186} = 2.54.$$

Then, determining if you should reject the null or not is as easy as determining if the obtained *t* value (+2.54) is in the predicted direction and if it is farther from zero than the critical value of +1.7613. In this case, the obtained *t* value is positive, as predicted, and it is more extreme than the critical value so you have sufficient evidence to reject the null hypothesis.

Another way to describe these results is to say, "The obtained *t* value is in the critical region; therefore, there is sufficient evidence to reject the null hypothesis." In this case, your $\alpha = .05$, and your obtained *t* were in the critical region. This means that there is less than a 5% chance that you would obtain a *t* value of +2.54 or higher if the null hypothesis were true. Therefore, the statistical analysis suggests that the null hypothesis is probably false and that the research hypothesis is likely to be true. In other words, the study suggests that psychology majors tend to overestimate their starting salaries after graduation.

Reading Question

17. If the obtained t value is farther from zero than the critical value, you should
    a. reject the null hypothesis.
    b. fail to reject the null hypothesis.

Step 5: Compute an Effect Size and Describe It

As with the z for a sample mean, after completing the significance test in Step 4, you should compute an effect size so you can describe the difference between the means as slight, moderate, or large. The effect size estimate for the single-sample t test is computed in the same way it was for the z for a sample mean except the sample standard deviation (i.e., SD) is the denominator rather than the population standard deviation (i.e., σ).

$$d = \frac{\text{Observed deviation between the means}}{\text{Standard deviation}} = \frac{M - \mu}{SD} = \frac{3{,}680 - 3{,}000}{1{,}038.68} = .65.$$

An effect size of .65 means that psychology majors estimated their mean salary to be .65 of a standard deviation higher than the actual average monthly salary of psychology majors. The same effect size guidelines are used for z and t—namely, ds close to .2, .5, and .8 are small, medium, and large effects, respectively. In this case, the effect size of .65 is medium to large, meaning that psychology majors overestimate their starting salaries by a quite a bit. However, with such a small sample size, it is possible that this medium to large effect is the result of sampling error.

Reading Question

18. When computing an effect size for a single-sample t test, the denominator is
    a. the sample standard deviation.
    b. the population standard deviation.

Step 6: Interpreting the Results of the Hypothesis Test

As before with the z for the sample mean, the final step is summarizing the results by explicitly stating if psychology majors tend to overestimate their starting salaries or not. Answering this question was the reason why you did all of these statistics; it's a waste of all that effort if you don't explicitly tell people what you found. Your summary should also include the mean of the sample and population, the df, the obtained t value, if the p value was less than or greater than the alpha value, and the computed effect size. Psychologists report all of this statistical information in a very specific format to save space. An example APA summary statement is shown as follows:

The average expected monthly salary for psychology majors (M = 680, SD = 1,038.68) was significantly higher than the actual starting salary of $3,000, $t(14) = 2.54$, $p < .05$, $d = .65$. This analysis suggests that psychology majors do tend to overestimate the starting salaries they can expect after graduation by quite a bit.

In the previous chapter, you learned to report the exact p value in your summary statement of a z for a sample mean (e.g., $p = .02$). *If you know the exact* p *value of a statistical result, you should always report it.* When using a z for a sample mean in the previous chapter, you could easily determine the exact p value for any z score by looking it up in the unit normal table (i.e., Appendix A). However, it is more difficult to determine the exact p value of a t score because while there is only one probability for each z score, the probability of a t score depends on the study's sample size. Therefore, it is not practical for any book to provide all the probabilities for every t score and every sample size. Therefore, it is common

to simply report if the *p* value is greater than or less than the chosen alpha value (e.g., .05 or .01) when performing a *t* test by hand. If the obtained *t* value is in the critical region, the *p* value for that *t* score must be less than the alpha value (.05 or .01). Therefore, you would write $p < .05$ whenever you reject the null. On the other hand, if the obtained *t* score is less than the critical value, its *p* value must be greater than the alpha value and you would write $p > .05$. To be clear, the only reason you would ever write $p < .05$ or $p > .05$ would be if you don't know the exact *p* value. When you compute a statistic by hand, you will not usually know the exact *p* value of your results. However, when you use a computer to perform your statistics, it will usually provide you with the exact *p* value and so you should report it.

Reading Question

19. The number in the parentheses in the character string "$t(14) = 2.22$, $p < .05$, $d = .57$" is the
    a. number of participants in the study.
    b. degrees of freedom.
    c. critical *t* value.

Reading Question

20. Generally, you should report _____, but it is acceptable to report _____ when you are computing your statistical tests by hand.
    a. $p > .05$ or $p < .05$; exact *p* values
    b. exact *p* values; $p > .05$ or $p < .05$

TWO-TAILED SINGLE-SAMPLE *t* EXAMPLE

In the previous example, we used a one-tailed *t* test (also known as a **directional test**). It is also possible to use a two-tailed *t* test (also known as a **nondirectional test**). For example, suppose that after learning that psychology majors were overly optimistic about their own starting salaries, you want to know if students are similarly optimistic when estimating others' starting salaries. To test this, you recruit 15 different psychology majors and ask them to estimate the average monthly salary that the typical psychology major can expect in his or her first job after graduation. You are not sure if students will overestimate or underestimate others' staring salaries, and so you should do a two-tailed test rather than a one-tailed test.

In two-tailed hypothesis tests, you do not have a specific directional prediction concerning the treatment. Therefore, the two-tailed research hypothesis states that the average estimated monthly salary will either be higher than $3,000 or lower than $3,000. Thus, two-tailed research hypotheses have two critical regions, one in the positive tail and one in the negative tail. If the obtained *t* value were in the positive critical region, you would conclude that the students overestimated the average starting salary of the typical psychology major. Conversely, if the obtained *t* value were in the negative critical region, you would conclude that the students underestimated the average starting salary. Finally, if the obtained *t* value were "close to zero" (i.e., between the positive and negative critical values), you would fail to reject the null hypothesis.

The choice between a one- and two-tailed test can have a dramatic impact on the critical region(s), and therefore, your decision about whether to use a one- or two-tailed test can influence whether or not you reject the null hypothesis. For example, if your sample has 15 people and you conduct a one-tailed test with $\alpha = .05$, your critical value would be +1.7613 and you would reject the null if your obtained *t* value was +1.8. However, if you analyzed the same data but chose to conduct a two-tailed test, your critical *t* values would be +2.1448 and −2.1448. If using this two-tailed test, you would not reject the null if your obtained *t* value was +1.8 because +1.8 is not in either of the two critical regions. Figure 7.2 makes this point visually. The reason for the different conclusions is that the one-tailed test critical value is closer to zero than the two-tailed critical values. For

both types of tests, 5% of the distribution is in a critical region, but with a two-tailed test, 2.5% of the distribution is in each tail while with a one-tailed test, 5% is in a single tail. This means that the critical value will be closer to zero for a one-tailed test than a two-tailed test. Therefore, if your research hypothesis is directional, it is to your advantage to use a one-tailed test rather than a two-tailed test. Again, this point is made visually in Figure 7.3.

Reading Question

21. Why is your decision to conduct a one-tailed rather than a two-tailed test potentially important?

a. One- and two-tailed tests have different critical regions and therefore may lead to different conclusions about the null hypothesis.

b. One- and two-tailed tests will produce different obtained *t* values.

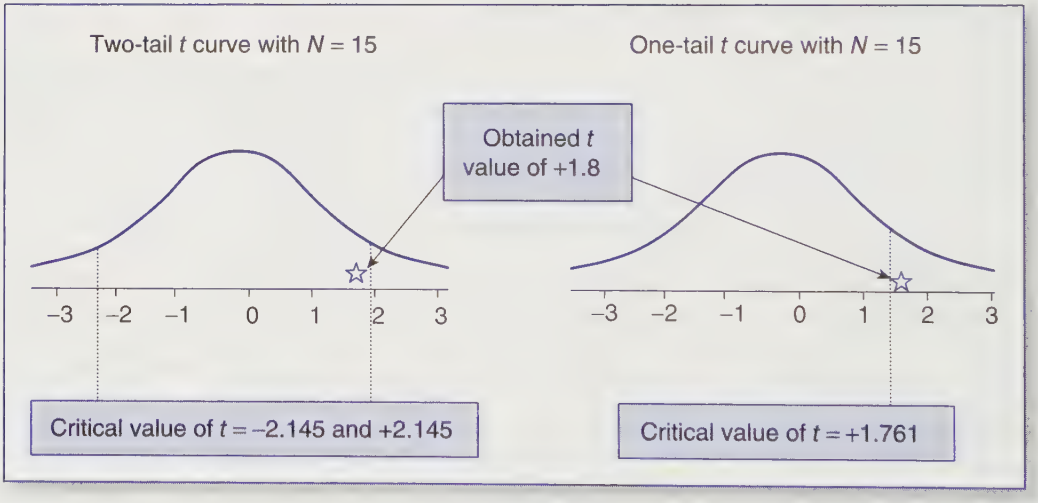
Reading Question

22. When should you conduct a one-tailed test?

a. Conduct a one-tailed test when the research hypothesis predicts scores will increase or if it predicts scores will decrease.

b. Conduct a one-tailed test when the research hypothesis does not specify if scores will increase or decrease.

Figure 7.3 Visual Representation of the Two-Tailed and One-Tailed Critical Regions



After learning that psychology majors were overly optimistic about their own starting salaries, you want to know if students are similarly optimistic about others' starting salaries. You recruit 15 different psychology majors and ask them to estimate the average monthly salary that the typical psychology major can expect in his or her first job after graduation. The estimated salaries were 3,400, 3,500, 2,500, 3,200, 2,700, 2,000, 2,500, 3,300, 3,200, 3,000, 2,900, 3,900, 3,800, 3,900, and 3,600. You know that the population's actual monthly salary is normally shaped with $\mu = 3{,}000$. You are not sure if

students will be optimistic about others' staring salaries, and so you correctly choose to do a two-tailed test rather than a one-tailed test. You decide to use $\alpha = .05$.

Step 1: Examine the Statistical Assumptions

Your study seems to meet all four statistical assumptions. You collected data carefully to ensure that the students' answers did not influence each other (*assumption of data independence*). You also worded your question carefully so students' answers were not made more or less variable (*assumption of homogeneity of variance*). You measured the DV of expected salary on an interval/ratio scale of measurement, and your IV identifies how the sample differs from the population (*appropriate measurement of variables assumption*). The distribution of sample means will have a normal shape (*assumption of normality*) because the original population of scores has a normal shape. Again, good experimental design is necessary for trustworthy statistics.

Step 2: State the Null and Research Hypotheses Symbolically and Verbally

The hypotheses for two-tailed tests are different from one-tailed tests. For a two-tailed test, the *null hypothesis* is always that the sample and population means will be the same. The *research hypothesis* is that the means will differ, but it does not indicate which mean will be higher than the other. The symbolic notation and the verbal descriptions of the null and research hypotheses for this study are in Table 7.3.

Table 7.3 Symbolic and Verbal Representations for Two-Tailed Research and Null Hypotheses for a Single-Sample *t* Test

	Symbolic	*Verbal*	*Mean Difference Created by*
Research hypothesis (H_1)	H_1: $\mu_{\text{psych majors}} \neq 3{,}000$	The mean estimated starting salary of the population of psychology majors *is not equal to* the actual mean starting salary ($3,000/month).	Psychology majors' over- or underestimate starting salaries of other psychology majors
Null hypothesis (H_0)	H_0: $\mu_{\text{psych majors}} = 3{,}000$	The mean estimated starting salary of the population of psychology majors *is equal to* the actual mean starting salary ($3,000/month).	Sampling error

Always use the "not equal" sign for a two-tailed research hypothesis. This indicates that you are predicting that the sample mean will be different from the population mean. You are not specifying whether the sample mean will be higher or lower than the population mean, just that it will be different.

Reading Question

23. Which of the following symbols should be used to represent a two-tailed research hypothesis?

a. ≤

b. >

c. =

d. ≠

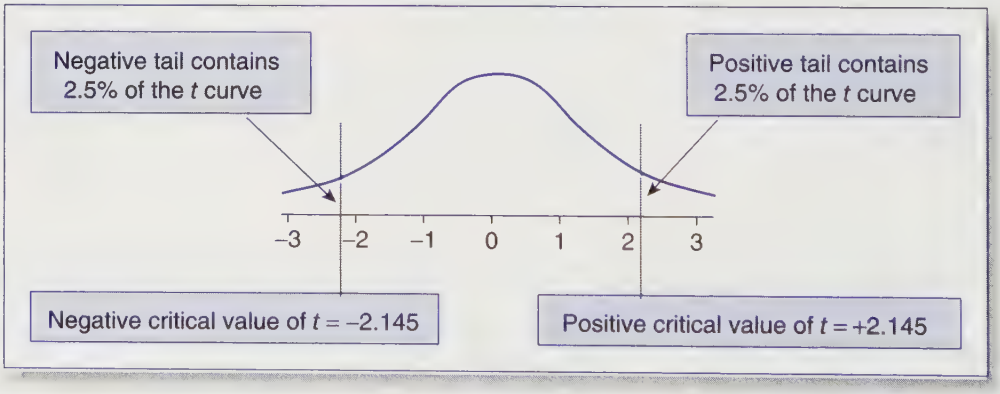
Step 3: Use Sample Size to Compute the Degrees of Freedom and Define the Critical Regions

The formula for *df* is the same for one- and two-tailed tests. As was the case in the one-tailed example, the *df* here is

$$df = N - 1 = 15 - 1 = 14.$$

Now you will use the table of critical values of *t* in Appendix B to determine the critical value for a two-tailed test when the $df = 14$. In this case, you find the *df* in the left-most column and then find the value in that row under the two-tailed .05 heading. The critical value is 2.1448. Therefore, the two cut lines that define the positive and negative critical regions are located at +2.1448 and −2.1448, respectively. Figure 7.4 will help you visualize these two critical regions on the distribution of *t* values.

Figure 7.4 Visual Representation of the Two Critical Regions



Step 4: Compute the Test Statistic (Single-Sample *t* Test)

4a. Compute the Deviation Between the Sample Mean and the Population Mean

This step is identical to a one-tailed test:

$$(M - \mu) = (3{,}160 - 3{,}000) = 160.$$

4b. Compute the Average Sampling Error Expected

This step is identical to a one-tailed test. As in the previous example, the standard deviation of the sample (i.e., *SD*) is not given to you in the problem, so you must compute it from the sample data. If you are not sure how this is done, you should review the previous example. In this problem, the standard error of the mean is defined by

$$SEM_s = \frac{SD}{\sqrt{N}} = \frac{561.630}{\sqrt{15}} = 145.012.$$

In this case, the typical amount of sampling error was computed to be 145.012. The correct way to think about this sampling error value is that while some sample means will have more sampling error than 145.012 and others will have less, the typical amount of sampling error for all possible samples is 145.012.

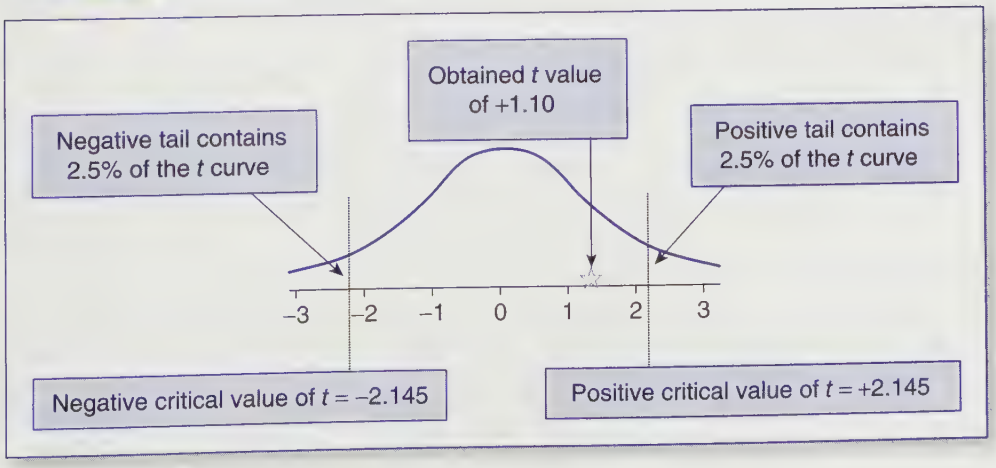
4c. Compute the Test Statistics (Single-Sample *t* Test)

The computation of the test statistic is also identical for one- and two-tailed tests.

$$t = \frac{M - \mu}{SEM_s} = \frac{3{,}160 - 3{,}000}{145.012} = 1.10.$$

The obtained *t* value associated with the sample mean of 3,160 (i.e., 1.10) was not in either critical region. This means that there is more than a 5% chance that the mean difference in the numerator is due to sampling error. Therefore, you should not reject the null hypothesis. The results suggest that the psychology majors do not overestimate the starting salaries of others. Figure 7.5 may help you visualize the outcome of this two-tailed hypothesis test. Because the obtained *t* value was not in either critical region, there is not enough evidence to reject the null hypothesis.

Figure 7.5 Visual Representation of the Relationship Between the Obtained *t* Value and the Two Critical Regions



Step 5: Compute an Effect Size and Describe It

The computation of the effect size is identical for one- and two-tailed tests.

$$d = \frac{\text{Observed deviation between the means}}{\text{Standard deviation}} = \frac{M - \mu}{SD} = \frac{3{,}160 - 3{,}000}{561.630} = 0.28.$$

The effect size of 0.28 indicates that psychology majors only slightly overestimate the starting salaries of others. Psychology students tend to overestimate the size of others' starting salaries by .28 of a standard deviation. The size if this effect is small, and because the sample size was very small, the effect size could very well be the result of sampling error.

Step 6: Interpreting the Results of the Hypothesis Test

The only difference in writing the summaries of one- and two-tailed hypothesis tests is that you need to let the reader know whether you ran a one- or two-tailed test. It is good practice to always do this, but we have not done so previously because you did not yet know about both types of tests. The following sentences summarize the analysis from this example:

The average estimated salary for typical psychology students (M = 3160.00, SD = 561.63) was not significantly different from the actual average monthly salary of $3,000, $t(14) = 1.10$, $p > .05$ (two-tailed), $d = .28$. These results suggest that psychology students are fairly accurate at estimating the future starting salaries of others. While there was a slight tendency to overestimate salaries, there was not sufficient evidence to conclude that the trend was more than just sampling error.

Reading Question

24. For which of the following hypothesis testing steps is there a difference between one-tailed and two-tailed tests? (You should choose three of the following five steps.)
    a. Step 2: stating null and research hypotheses
    b. Step 3: defining the critical region
    c. Step 4: computing the test statistic
    d. Step 5: computing the effect size
    e. Step 6: writing the results

OTHER ALPHA LEVELS

In both of the preceding examples, we chose to use $\alpha = .05$ because in many research areas, it is the most commonly used alpha level. However, in some cases, you may choose the more stringent alpha level of .01. One result of using $\alpha = .01$ is that the critical region becomes smaller. For example, if you were to use $\alpha = .01$ with a two-tailed test with $df = 14$, your critical regions would be $t \geq 2.9768$ and $t \leq -2.9768$. If you were to use $\alpha = .01$ with a one-tailed test with $df = 14$, your critical regions would be $t \geq 2.6245$.

When the null hypothesis is true, the probability of a Type I error is equal to the alpha value you choose. Therefore, if you choose the more stringent $\alpha = .01$, you are making it harder to reject the null hypothesis. This means that you are *less likely to make a Type I error,* but the smaller alpha level also affects the Type II error rates and statistical power. The smaller alpha of .01 means you have less statistical power and are *more likely to make a Type II error* than if you use a .05 alpha value. Thus, you need to balance your desire to minimize Type I errors with your desire to have statistical power (i.e., your ability to reject the null hypothesis).

In this book, we will typically tell you which alpha level to use because you don't yet have enough information about the research questions to make an informed decision about which alpha level is most appropriate. Researchers make this decision based on a number of different factors. For example, a researcher who is testing a treatment for a serious disorder that currently has no available treatment is going to be most concerned about making a Type II error (missing a treatment really works) and use a higher alpha level (e.g., .05 or even .10). Researchers may also choose their alpha level based on their sample size. For example, a researcher studying gender differences in math test scores may have access to scores from millions of children. In this case, the researcher would probably use a much lower alpha level (e.g., .01 or even .001) because they have so much statistical power that they don't want to report effects that are statistically significant but practically meaningless (i.e., have very small effect sizes). Choosing the correct alpha value is highly dependent on the research context.

Reading Question

25. Which of the following alpha values creates a greater probability of a Type I error?
    a. .01
    b. .05

Reading Question

26. Which of the following alpha values creates a greater probability of a Type II error?
    a. .01
    b. .05

Reading Question

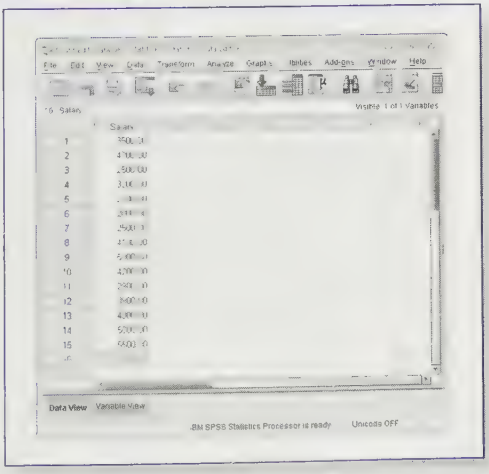
27. Which of the following alpha values results in more statistical power?
    a. .01
    b. .05

SPSS

Data File

Enter the scores for the first problem (students' expected salaries) as described in Chapter 1. When you are done, your data file should look like Figure 7.6.

Figure 7.6 SPSS Screenshot of Data Entry Screen



Computing a Single-Sample *t*

- Click on the Analyze menu. Choose Compare Means and then One Sample *t* Test (see Figure 7.7).
- Move the variable of interest into the Variables box (see Figure 7.8).
- Change the Test Value if necessary (in this case it should be 3000).
    - The Test Value is the number you are comparing the sample mean to.
    - If your *t* is wrong, it is most likely because you left the Test Value as 0.
- Click on the OK button.

Figure 7.7 SPSS Screenshot of Selecting the One-Sample *t* Test

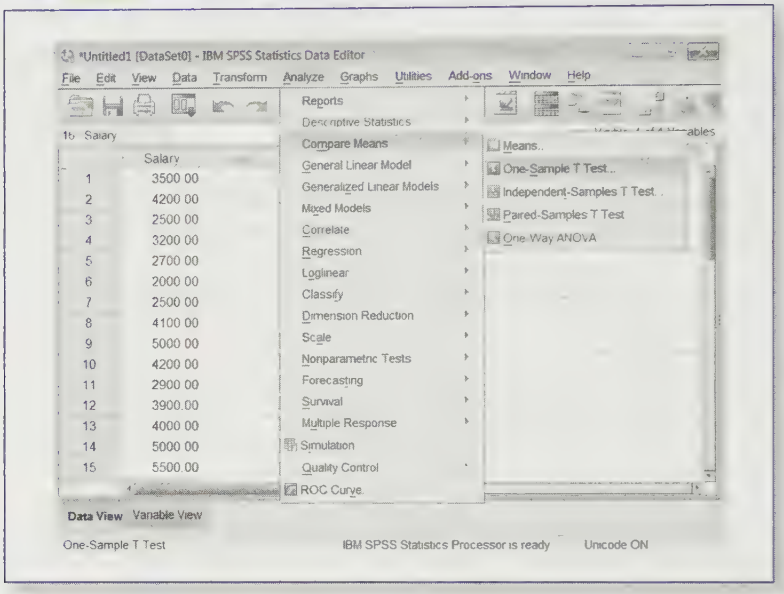
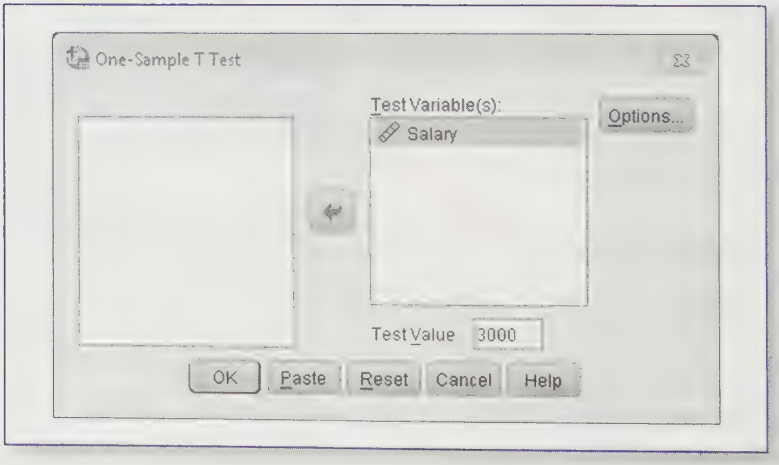


Figure 7.8 SPSS Screenshot of the Test Value and Test Variable for the One-Sample *t* Test



Output

Your output will have the following two boxes (see Figure 7.9).

Figure 7.9 Annotated SPSS Output for a One-Sample *t* Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Salary	15	3680.0000	1038.68048	268.18615

N: Sample size

Mean: Sample mean (*M*)

Std. Deviation: Sample standard deviation (*SD*)

Std. Error Mean: The average distance that all possible sample means of 15 people are from the population mean of 100. This is the estimated sampling error and the denominator of *t* test (SEM_s).

Test Value: Population mean (μ) or the value to which the sample mean is compared

One-Sample Test

	Test Value = 3000					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Salary	2.536	14	.024	680.00000	104.7979	1255.2021

t: Obtained *t* value

df: Degrees of freedom (*N* - 1)

Sig. (2-tailed): Two-tailed *p* value; the probability of obtaining a *t* score this extreme or more extreme if the null hypothesis is true.

for a one-tailed test, divide *p* by 2;

reject H_0 if $p < \alpha$

Mean Difference: The difference between the sample mean and the population mean; the numerator of the *t* test

95% Confidence Interval: We are 95% confident that the actual difference between the sample and population means is between the lower and upper values.

Reading Question

28. When you enter data into SPSS, each person's data (i.e., score) go in their own

a. row.

b. column.

Reading Question

29. Use the "One-Sample Statistics" output to determine how many people were in this study (i.e., the size of the sample).
    a. 14
    b. 15

Reading Question

30. Use the "One-Sample Statistics" output to determine what the mean of the sample was in this study.
    a. 3,000
    b. 3,680
    c. 1,038.68
    d. 680

Reading Question

31. Use the "One-Sample *t* test" output to determine the population mean (i.e., the test value) in this study.
    a. 3,000
    b. 3,680
    c. 1,038.68
    d. 680

Reading Question

32. Use the "One-Sample Test" output to determine the obtained *t* value in this study.
    a. .024
    b. 2.536
    c. 680

Reading Question

33. Use the "One-Sample Test" output to determine the *p* value in this study.
    a. .024
    b. .05
    c. 2.536

Reading Question

34. Use the "One-Sample Test" output to determine if the null hypothesis of this study should be rejected.
    a. The null hypothesis should be rejected because the *p* value (Sig.) is less than .05.
    b. The null hypothesis should not be rejected because the *p* value (Sig.) is less than .05.

OVERVIEW OF THE ACTIVITY

In Activity 7.1, you will work through all of the steps of hypothesis testing doing the calculations with a calculator and using SPSS. You will also read research scenarios and determine whether a *z* for a sample mean or a single-sample *t* test is the correct test for that situation.

Activity 7.1: Single-Sample *t* Test

Learning Objectives

After completing this activity, you should be able to do the following:

- Write a null hypothesis and research hypothesis using symbols and words
- Compute the degrees of freedom and define a critical region
- Compute a single-sample *t* using a calculator and SPSS
- Determine whether you should reject the null hypothesis
- Compute and interpret an effect size (*d*)
- Summarize the results of the analysis using APA style
- Interpret the SPSS output for a single-sample *t*
- Explain how you decide to reject the null hypothesis based on the Sig. (*p*) value provided by SPSS for both one- and two-tailed tests

TWO-TAILED EXAMPLE

Renner and Mackin (1998) gave a survey assessing levels of stress to 239 students at a university in the eastern United States. The survey measures stress on an interval/ratio scale. The mean stress score for this group was normally distributed with a mean of 1,247. The Women's College Coalition argues that students at women's colleges are more satisfied with their college experience than students at coed colleges. You wonder if those attending a women's college have different stress levels than those attending coed colleges. You gave the same scale used by Renner and Mackin (1998) to a sample of 12 students from a women's college to determine if their stress levels were *significantly different* from the students in the original Renner and Mackin (1998) study. The data from the 12 students are as follows:

200	900	740
850	500	400
2,300	650	700
1,300	950	1,000

1. All the necessary assumptions are met by this study. Match each of the statistical assumptions to the fact that implies the assumption was met.

   d Independence

   b Appropriate measurement of the IV and the DV

   c Normality

   a Homogeneity of variance

a. Based on how the data were collected, it is unlikely that the *SD* and the σ will be dramatically different.

b. The IV identifies how the sample is distinct from the population, and the DV is measured on an interval/ratio scale.

c. The population of stress scores has a normal shape.

d. The stress scores from each person in the sample were collected in a manner that ensured one person's scores did not affect anyone else's scores.

2. Why can't you use a *z* test to analyze these data?

a. You don't have a sample mean and a standard deviation.

b. You don't have a population standard deviation.

Analyze the data using a calculator. *Use a two-tailed test with an alpha of .05* to determine if the mean score for this sample is significantly different from a score of 1,247 (the mean score produced by the normative sample).

3. Choose the correct two-tailed null hypothesis.

a. $\mu_{\text{women's college}} = 1{,}247$

b. $\mu_{\text{women's college}} \neq 1{,}247$

c. $\mu_{\text{women's college}} > 1{,}247$

d. $\mu_{\text{women's college}} \leq 1{,}247$

4. Choose the correct two-tailed research hypothesis.

a. $\mu_{\text{women's college}} = 1{,}247$

b. $\mu_{\text{women's college}} \neq 1{,}247$

c. $\mu_{\text{women's college}} > 1{,}247$

d. $\mu_{\text{women's college}} \leq 1{,}247$

5. Compute the degrees of freedom ($df = N - 1$) and use that value to define the critical regions Use $\alpha = .05$.

a. Reject H_0 if $t > 1.96$ or $t < -1.96$

b. Reject H_0 if $t > 2.2010$ or $t < -2.2010$

c. Reject H_0 if $t < 1.96$

d. Reject H_0 if $t < 2.2010$

6. Compute the test statistic. (Note: You will need to compute the sample *M* and *SD* using the 12 scores from the sample.)

7. Compute an effect size (*d*).

8. Describe the size of the effect as small, small to medium, medium, medium to large, or large.

9. Which of the following is the best interpretation of the effect size?
   a. The sample's mean stress level was below the population's mean stress level by 0.70 points.
   b. The sample's mean stress level was 0.70 standard deviations below the population mean's stress level.
   c. The small effect size indicates that the mean difference was not statistically significant.
   d. The effect size (0.70) was greater than .05 and so we fail to reject the null.

10. Choose one of the following APA style summaries for the analysis you just completed. Fill in the three blanks with the (*df*), obtained *t* value, and the effect size (i.e., *d*) *only for the correct summary.*
    a. The students' scores ($M = 874.17$, $SD = 535.56$) were significantly lower than the population ($\mu = 1{,}247$), t(_____) = _____, $p < .05$ (two-tailed), d = _____.
    b. The students' scores ($M = 874.17$, $SD = 535.56$) were significantly higher than the population ($\mu = 1{,}247$), t(_____) = _____, $p < .05$ (two-tailed), d = _____.
    c. The students' scores ($\mu = 874.17$, $SD = 535.56$) were not significantly different from the population ($M = 1{,}247$), t(_____) = _____, $p < .05$ (two-tailed), d = _____.

11. You may have made an error when you decided whether to reject the null. Based on your answer to Question 10, which error is possible (Type I or Type II)?
    a. It is possible that I made a Type I error because I rejected the null hypothesis.
    b. It is possible that I made a Type II error because I rejected the null hypothesis.
    c. It is possible that I made a Type I error because I failed to reject the null hypothesis.
    d. It is possible that I made a Type II error because I failed to reject the null hypothesis.

12. Use SPSS to reanalyze the data you just analyzed by hand.
    - Enter the data into one column in the SPSS data editor.
    - Click on the Analyze menu. Choose Compare Means and then One-Sample *t* Test.
    - Move the Stress Scores variable into the Variables box.
    - Change the Test Value to the μ of the comparison group (i.e., 1247).
    - Click on the OK button.

Locate each number that you computed by hand in the SPSS output. It may help to print the output and label it as you find each value. What information did you need when you computed this problem by hand that you do not need when you use SPSS?

   a. The critical value
   b. The sample mean
   c. The population standard deviation

13. How do you decide if you should reject the null hypothesis when doing *hand computations?*
    a. You reject the null hypothesis if the obtained *t* value is in the critical region.
    b. You reject the null hypothesis if the obtained *t* value is outside of the critical region.

14. How do you decide if you should reject the null hypothesis when *using SPSS* and doing a two-tailed test?
    a. Divide the Sig. value by 2 and then you reject the null if the Sig. value is less than the alpha.
    b. Divide the Sig. value by 2 and then you reject the null if the Sig. value is greater than the alpha.
    c. You reject the null if the Sig. value is less than the alpha.
    d. You reject the null if the Sig. value is greater than the alpha.

15. How do you decide if you should reject the null hypothesis when *using SPSS* and doing a one-tailed test?
    a. Divide the Sig. value by 2 and then you reject the null if the Sig. value is less than the alpha.
    b. Divide the Sig. value by 2 and then you reject the null if the Sig. value is greater than the alpha.
    c. You reject the null if the Sig. value is less than the alpha.
    d. You reject the null if the Sig. value is greater than the alpha.

ONE-TAILED EXAMPLE

The Women's College Coalition argues that students at women's colleges are more satisfied with their college experience than students at coed colleges. You wonder if this is related to stress. Specifically, you want to determine if students at a women's college experience *less* stress than the general population does. To test this hypothesis, you give the stress survey to $N = 1{,}150$ students at a women's college and obtain a mean score of $M = 1{,}198$ with a standard deviation of $SD = 654$. Are the female college students *less* stressed than the general population, $\mu = 1{,}247$?

16. Match each of the statistical assumptions to the fact that implies that assumption is met.
    c Independence
    b Appropriate measurement of the IV and the DV
    a Normality
    d Homogeneity of variance

    a. The sample size is $N = 1{,}150$.
    b. The grouping variable identifies a difference between the sample and the population.
    c. The data were collected so individuals' responses did not influence the responses of others.
    d. There is no reason to suspect dramatically difference standard deviations for the sample and the population.

17. Choose the correct null hypothesis.
    a. The population of students at the women's college is less stressed than the general population.
    b. The population of students at the women's college is not less stressed than the general population.

c. The students at the women's college will have stress levels that are not significantly different from the general population.

d. The students at the women's college will have stress levels that are significantly different from the general population.

18. Choose the correct research hypothesis.

a. The students at the women's college will be less stressed than the general population.

b. The students at the women's college will not be less stressed than the general population.

c. The students at the women's college will have stress levels that are not significantly different from the general population.

d. The students at the women's college will have stress levels that are significantly different from the general population.

19. Why should you use a one-tailed significance test for this research scenario?

a. The sample mean is lower than the population mean.

b. The population standard deviation is not provided.

c. The researcher has a clear prediction about the direction of the results.

d. The sample size is sufficiently large.

20. Compute the degrees of freedom and use that value to define the critical region for a one-tailed test with an alpha of .05.

a. Reject H_0 if $t > 1.96$ or $t < -1.96$.

b. Reject H_0 if $t > 1.645$ or $t < -1.645$.

c. Reject H_0 if $t < -1.96$.

d. Reject H_0 if $t < -1.645$.

21. What effect does increasing the sample size have on the critical value?

a. It raises the critical value (it moves farther from zero).

b. It lowers the critical value (it moves closer to zero).

22. Compute the test statistic.

23. Compute an effect size (*d*).

24. Describe the size of the effect as small, small to medium, medium, medium to large, or large.

25. Write an APA-style summary of the results. Do more than say reject or fail to reject the null hypothesis. Explain what this means in terms of the independent variable (IV) and the dependent variable (DV). Use the options provided in Question 10 as a guide.

26. Does this analysis provide strong evidence that students at women's colleges experience less stress than the general population?
    a. Yes, the results were statistically significant. The effect size is only something that you need to consider if the sample size is small and/or the results were not statistically significant.
    b. Yes, we rejected the null, which indicates that the result was significant. The effect size was small, but we can have a high degree of confidence in the effect size because the sample size was quite large.
    c. No, the results were not statistically significant and the effect size was very small.
    d. No, although we rejected the null hypothesis, the effect size was quite small. This suggests that the difference is not practically important.

27. In general, a larger sample size results in a _____ denominator of the *t* statistic, which will _____the value of the obtained *t*.
    a. smaller; increase
    b. smaller; decrease
    c. larger; increase
    d. larger; decrease

28. The sample sizes varied widely in the two studies you analyzed in this activity. This change in sample size changed the critical values. Why does increasing the sample size change the critical value for a *t* test but not a *z* test? Select all that apply.
    a. The distribution of sample means changes shape (becomes more bell shaped) as the sample size increases for the *t* distribution but not for the *z* distribution.
    b. The *t* distribution is always the same shape regardless of sample size, but the *z* distribution changes shape as sample size increases.
    c. For the *t* distribution, as the sample size increases, the sample standard deviation gets closer to the population standard deviation and the distribution looks more like the *z* distribution.

29. True or false: The critical value for a *t* test changes with sample size, but the size of the critical region does not change; it is still equal to the alpha level.

30. When computing the single-sample *t*, we use the sample standard deviation rather than the population standard deviation to compute the test statistic. Why would it be better to use the population standard deviation if possible? Select all that apply.
    a. The sample standard deviation is only an estimate of the population parameter.
    b. You are more likely to reject the null hypothesis with a *z* test than with a *t* test.
    c. The population standard deviation is more accurate than the sample standard deviation.

CHOOSE THE CORRECT STATISTIC

From this point forward, each chapter of this book introduces a new statistic that can be used in different research situations. In each chapter, it is obvious which statistic should be used because we work with only one statistic in each chapter. However, when you collect your own data, you must determine which statistic

should be used to answer your research question. Right now, you know two different statistics: (1) the z for a sample mean and (2) the single-sample t.

Both the z for a sample mean and a single-sample t determine if a sample mean is significantly different from a population mean or some value of theoretical interest. The primary difference between these significance tests is that to compute the z for a sample mean, you must know the population standard deviation (σ). You do not need to know the population standard deviation when using a single-sample t.

Determine which statistic should be used in each of the following research scenarios: z for sample mean or single-sample t.

31. A researcher wants to compare a sample mean, to a population mean, but the population standard deviation is unknown.

32. A researcher wants to compare a sample mean, to a population mean, and the population standard deviation is known.

34. A researcher wants to compare a sample mean to a value of theoretical interest.

35. New Army recruits are required to take a test called the Armed Services Vocational Aptitude Battery (ASVAB). This test measures vocational aptitude on an interval/ratio scale and is standardized so that it is normally distributed with a mean of $\mu = 50$ and a standard deviation of $\sigma = 10$. A recruiter wonders if soldiers who are nurses have higher than average ASVAB scores. To test this, the recruiter records the ASVAB scores of a sample of 37 nurses and finds that their mean score was 58 ($SD = 12$). Is this significantly higher than 50?

36. An industrial/organizational psychologist wonders if nurses in the army are more or less satisfied with their career as nurses than nurses in the civilian sector. Previous research conducted by the psychologist revealed that civilian nurses rated their overall job satisfaction as a 7.3 on a 10-point scale. This measure of job satisfaction is an interval/ratio scale. To determine if army nurses have different levels of job satisfaction, the psychologist recruited a sample of 58 nurses in the army and asked them to indicate their overall level of job satisfaction on a 10-point scale with 1 = *not at all satisfied* and 10 = *very satisfied*. The mean for the sample was 7.9 with a standard deviation of 1.64.

37. In an attempt to boost sales, the manager of a grocery store orders carts that are 25% larger than carts currently used in the grocery store. He reasons that people will be more likely to purchase more to fill up their carts. For the year prior to purchase of the new carts, the average amount spent for each shopper was normally distributed with $\mu = \$125.56$ ($\sigma = 38.98$). The average amount spent for a sample of 152 shoppers was $M = \$142.11$ ($SD = \$43.21$). Did the carts lead to a significant increase in sales?

Chapter 7 Practice Test

1. How is the single-sample *t* different from the *z* for a sample mean?
   a. The *z* for a sample mean requires a larger sample size than the single-sample *t*.
   b. The *z* for a sample mean requires knowledge of a population standard deviation, and the single-sample *t* does not.
   c. The single-sample *t* requires knowledge of a population mean, and the *z* for a sample mean does not.
   d. The *z* for a sample mean allows you to compute a Type I error rate, but the single-sample *t* does not allow for these calculations.

2. As sample size increases, what happens to the estimated standard error of the mean?
   a. It increases.
   b. It decreases.

3. What does the estimated standard error measure?
   a. The typical distance between the *t* scores in the distribution
   b. The typical distance between the all possible sample means of a given size and the population mean
   c. The typical distance between the scores in the population and the mean of the population

4. Which of the following will increase as the sample size increases (assuming everything else is held constant)? Select two.
   a. Type I error rate
   b. Statistical power
   c. Sampling error
   d. Critical *t* value
   e. Size of the critical region
   f. Computed *t* value

5. The following results from a single-sample *t* are reported in a journal article, $t(59) = 2.24$, $p = .03$, $d = .78$. How many people participated in the study?
   a. 57
   b. 58
   c. 59
   d. 60

6. The following results from a single-sample *t* are reported in a journal article, $t(59) = 2.24$, $p = .03$, $d = .78$. Should the researcher reject or fail to reject the null hypothesis?
   a. Reject
   b. Fail to reject

7. The following results from a single-sample *t* are reported in a journal article, $t(59) = 2.24$, $p = .03$, $d = .78$. How would you describe the size of the effect?

a. Small
b. Small to medium
c. Medium
d. Medium to large
e. Large

8. Researchers can only make a Type I error if they
   a. reject the null hypothesis.
   b. fail to reject the null hypothesis.

9. Researchers can only make a Type II error if they
   a. reject the null hypothesis.
   b. fail to reject the null hypothesis.

10. A representative sample of Americans were asked to indicate their agreement with a series of eight statements about childhood vaccines, such as the following:

    Childhood vaccines should be required.

    Strongly Disagree 1 2 3 4 5 Strongly Agree

    The responses to the eight questions were averaged to form a single measure of support of childhood vaccines. The scores formed an interval scale that was normally distributed. The mean response on this scale for the general public was 3.7. A researcher wonders if scientists are more likely to agree with this statement than the public. To test this possibility, she obtains a sample of 36 scientists and asks them the same question. Their average level of agreement with this statement was 4.1 with a standard deviation of .9. Is the value for the sample of scientists significantly higher than 3.7?

    What is the one-tailed research hypothesis for this study?

    a. $\mu_{scientists} \geq 3.7$
    b. $\mu_{scientists} \geq 4.1$
    c. $\mu_{scientists} \geq 3.7$
    d. $\mu_{scientists} \geq 4.1$
    e. $\mu_{scientists} = 3.7$
    f. $\mu_{scientists} \neq 4.1$

11. What is the one-tailed null hypothesis for this study?
    a. $\mu_{scientists} \leq 3.7$
    b. $\mu_{scientists} \leq 4.1$
    c. $\mu_{scientists} \leq 3.7$
    d. $\mu_{scientists} \leq 4.1$
    e. $\mu_{scientists} = 3.7$
    f. $\mu_{scientists} \neq 4.1$

12. Should this researcher be concerned with violating any of the assumptions associated with this test?
    a. Yes, it violates the assumption of appropriate measurement. The DV is not measured on an interval/ratio scale.
    b. Yes, the sample size is too small to assume a normal distribution of sample means.
    c. Yes, it is very likely that the treatment will make the sample scores far less variable than the population.
    d. No, all assumptions are met. The researcher can conduct the hypothesis test.
13. Compute the degrees of freedom.
    a. 34
    b. 35
    c. 36
    d. 37
14. What is the critical t value for this one-tailed test with $\alpha = .05$?
    a. 1.6896
    b. 2.4377
    c. 1.6883
    d. 2.4645
15. When you locate the critical region, which distribution of sample means is the critical region on?
    a. The null distribution
    b. The research distribution
16. Compute the single-sample t for these data.
    a. .44
    b. .12
    c. 2.67
    d. 2.46
17. Should you reject or fail to reject the null hypothesis?
    a. Reject
    b. Fail to reject
18. The researcher might have made a _____ error.
    a. Type I
    b. Type II
19. What do the results of the significance test tell you?
    a. That the scientists are less likely to support mandated childhood vaccines than the public

b. That the difference observed between scientists' attitudes toward childhood vaccines and the public's attitudes is unlikely to be due to sampling error

c. How large the difference is between the scientists' attitudes toward childhood vaccines and the public's attitudes

d. The difference between the scientists' attitudes toward childhood vaccines and the public's attitudes in standard deviation units

20. Compute the effect size (*d*).

a. .44

b. .12

c. 2.67

d. 2.46

21. What does the effect size tell you? Select all that apply.

a. That the scientists are less likely to support mandated childhood vaccines than the public

b. That the difference observed between scientists' attitudes toward childhood vaccines and the public's attitudes is likely to be due to sampling error

c. How large the difference is between the scientists' attitudes toward childhood vaccines and the public's attitudes

d. The difference between the scientists' attitudes toward childhood vaccines and the public's attitudes in standard deviation units

22. After hearing of these results regarding the differences between the attitudes of scientists and the public, a researcher wonders if this difference is a function of level of education overall or if it is specific to science education. Thus, this researcher obtains a sample of 50 people who have a PhD in the humanities and plans to compare their mean level of agreement to the mean level of agreement of scientists (4.1). He is not sure what to expect, and so he chooses to do a two-tailed test. What is the two-tailed null hypothesis for this study?

a. $\mu_{humanities} \geq 4.1$

b. $\mu_{humanities} \geq 4.1$

c. $\mu_{humanities} = 4.1$

d. $\mu_{humanities} \neq 4.1$

23. What is the two-tailed research hypothesis for this study?

a. $\mu_{humanities} \geq 4.1$

b. $\mu_{humanities} \geq 4.1$

c. $\mu_{humanities} = 4.1$

d. $\mu_{humanities} \neq 4.1$

24. The SPSS output from the analysis follows:

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
VaccineAttitudes	50	3.8000	1.12486	.15908

One-Sample Test

	Test Value = 4.1					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
VaccineAttitudes	-1.886	49	.065	-.30000	-.6197	.0197

Should the researcher reject or fail to reject the null hypothesis?

a. Reject

b. Fail to reject

25. Which type of error might this researcher have made?

a. Type I

b. Type II

26. Compute the effect size (d).

a. .27

b. 1.135

c. 3.39

27. How large is this effect?

a. Small

b. Small to medium

c. Medium

d. Medium to large

e. Large

28. Choose the correct APA style summary of the results:

a. The average level of support for vaccines for people in the humanities with PhDs was $M = 3.80$, $SD = 1.12$. The average level of support for scientists was 4.1. This difference was not significant, $t(49) = -1.89$, $p = .07$, $d = -27$, and the effect size was small.

b. The average level of support for vaccines for people in the humanities with PhDs was $M = 3.80$, $SD = 1.12$. The average level of support for scientists was 4.1. This difference was not significant, $t(49) = -1.89$, $p = .07$, $d = -27$, and the effect size was small. Overall, scientists were more supportive of childhood vaccines than were those in the humanities.

c. The average level of support for vaccines for people in the humanities with PhDs was $M = 3.80$, $SD = 1.12$. The average level of support for scientists was 4.1. This difference was significant, $t(49) = -1.89$, $p = .07$, $d = -27$, and the effect size was small.

d. The average level of support for vaccines for people in the humanities with PhDs was $M = 3.80$, $SD = 1.12$. The average level of support for scientists was 4.1. This difference was significant, $t(49) = -1.89$, $p = .07$, $d = -27$, and the effect size was small. Overall, scientists were more supportive of childhood vaccines than were those in the humanities.

29. Which statistical procedure do you use to determine if a sample mean is significantly different than a population mean?
    a. Null hypothesis significance test
    b. Effect size

30. Which statistical procedure do you use to describe the size of the difference between a sample mean and a population mean?
    a. Null hypothesis significance test
    b. Effect size

31. Salaries at an organization are normally distributed with a mean of $\mu = \$50{,}000$ with a standard deviation of $\sigma = \$15{,}000$. What is the probability of selecting a sample of 40 people with a mean salary less than $40,000? Which statistic should you use to answer this question?
    a. *z* for a single score
    b. *z* for a sample mean
    c. Single-sample *t*

32. Scores on a test of emotional intelligence are normally distributed with a mean of 100 and a standard deviation of 12. What proportion of people have emotional intelligence scores that are more than two standard deviations above the mean? Which statistic should you use to answer this question?
    a. *z* for a single score
    b. *z* for a sample mean
    c. Single-sample *t*

33. Banks are bound to make the occasional error, but a customer advocacy group wonders if the errors are truly random. If they are random, the dollar amount of the errors should average out to zero because some of the errors would be in the bank's favor (positive amounts) and the other errors should be in the customers' favor (negative amounts). To investigate this possibility, the advocacy group obtains a list of the 100 most recently discovered errors. The average amount of the error is 34.85 (in the bank's favor) with a standard deviation of 15.32. Which statistic should you use to answer this question?
    a. *z* for a single score
    b. *z* for a sample mean
    c. Single-sample *t*

REFERENCES

Renner, M. J., & Mackin, R. (2002). A life stress instrument for classroom use. In R. A. Griggs (Ed.), *Handbook for Teaching Introductory Psychology: Vol. 3. With an emphasis on assessment* (pp. 236–238). Mahwah, NJ: Lawrence Erlbaum.

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology, 39*(5), 806–820.

CHAPTER 8

Estimation With Confidence Intervals

LEARNING OBJECTIVES

After reading this chapter, you should be able to do the following:

- Describe the distinct purposes of significance testing, effect sizes, and confidence intervals
- Explain the logic of confidence intervals
- Compute a confidence interval for a population mean
- Compute a confidence interval for a mean difference between a sample mean and a population mean
- Report confidence intervals in American Psychological Association (APA) style
- Identify correct and incorrect interpretations of confidence intervals

THREE STATISTICAL PROCEDURES WITH THREE DISTINCT PURPOSES

In Chapter 7, you learned how to use the single-sample *t* test to determine whether an observed difference between a sample mean and a population mean is likely to be created by sampling error or by an independent variable (IV; i.e., the grouping variable). The single-sample *t* test compares the mean difference observed in the data to the mean difference expected due to sampling error (i.e., *SEM*). If the observed mean difference is significantly greater than the mean difference expected due to sampling error, the null hypothesis is rejected and you conclude that the mean difference was created by the IV. This conclusion is sound if the study's research methodology is sound (i.e., it does not have confounding variables).

Reading Question

1. Significance tests (e.g., single-sample *t* tests) were designed to help researchers determine if the observed difference was likely to be created by
   a. sampling error.
   b. confounds.

After researchers use significance tests, they compute an effect size (e.g., *d*) to describe the magnitude of the IV's impact on the dependent variable (DV). Computing an effect size is just as important as significance testing. For example, you may use the result of a significance test to reject the null hypothesis and then find that the effect size is very small. This would indicate that although the results are unlikely to be due to sampling error, the size of the effect may be too small to be important or useful.

Reading Question

2. If a statistical test leads to rejecting the null hypothesis, the research finding is always important or useful.
   a. True
   b. False

Reading Question

3. An effect size can help a researcher determine
   a. if an effect is large enough to be useful.
   b. if the mean difference is likely to be due to sampling error.
   c. both of the above.

In this chapter, you will learn about a third type of statistical procedure that has a different purpose from either significance testing or effect size. This third type of statistical procedure, called a confidence interval, helps researchers generalize the results of their studies to an entire population. These three statistical procedures should be used together routinely. When the results of all three procedures are interpreted correctly, the combination constitutes a wealth of information that greatly facilities scientific reasoning.

For example, the provost of a college is considering adding a "healthy living" class to the general education requirements for students but first wants to determine if taking such a class increases healthy behaviors. Although the course will discuss a variety of healthy behaviors, the provost is especially concerned with exercise because the data clearly indicate that exercise is associated with a host of positive outcomes in both physical and mental health. Last year, the provost's office surveyed the entire population of students and found that they reported exercising an average of 150 minutes each week. To test the effectiveness of the healthy living class, 60 college students were recruited and all 60 took the new class. After completing the class, participants wore a heart rate monitor for 1 month, enabling researchers to accurately determine the average exercise of each participant in minutes per week. The provost opts to do a two-tailed test because although he hopes that the course will increase exercise behavior, he is concerned that people will respond negatively to the course and actually decrease the amount of time they spend exercising. A single-sample *t* test using $\alpha = .05$ revealed that the average number of minutes students who took the healthy living class exercised each week was $M = 155.67$ ($SD = 20.85$). Thus, there was a 5.67-minute difference between the students who took the healthy living course and the population mean of 150 minutes. The single-sample *t* test indicated that this difference was unlikely to be due to sampling error, $t(59) = 2.11$, $p = .04$ (two-tailed), $d = .27$.

Specifically, the *p* value ($p = .04$) indicates that only 4 out of 100 times would a mean difference of 5.67 (155.67 − 150) or larger occur due to sampling error when the null hypothesis is true. Therefore, the healthy living class probably created the mean difference. The effect size ($d = .27$) indicates that the impact of the class was small in size. Specifically, the *d* indicates that the class increased the mean exercise time by .27 of a standard deviation. Now, the provost is ready to generalize these results to the entire population of students at the college. In other words, he wants an estimate of what the mean exercise time of the entire population of students at the college would be *if they took the healthy living class*. The best estimate of this parameter is the mean of the sample that took the class—namely,

155.67 minutes. However, due to sampling error, this sample mean probably is not perfectly accurate. A *confidence interval* is a statistical procedure that uses the same expected sampling error formula (i.e., SEM_s) as hypothesis tests do, but in a different way, to create a *range of plausible values for the population parameter*. For example, in this situation, the researcher could use a confidence interval to conclude with 95% confidence that the mean number of minutes of exercise in the population *if they all took the healthy living course* would be between 150.29 and 161.05 minutes each week (you'll learn how to compute this confidence interval later in the chapter). He could also estimate the *mean difference* in exercise that would result if the population took the class. In this situation, he could conclude with 95% confidence that exercise time would increase between .29 minutes and 11.05 minutes each week if the population took the healthy living course. Both of these confidence intervals provide very important information to the researcher. For example, knowing that exercise times would be between 150.28 and 161.05 minutes a week *if the population took the class* might help the provost decide whether or not to make the class required for all college students. Similarly, knowing that requiring the class would increase the average weekly exercise time between .28 and 11.05 minutes might also help the provost consider the impact this change might have on students' lives.

As you can see, confidence intervals are an extremely useful statistical procedure. They help apply sample results to populations. In many situations, describing how research results generalize to a population is the ultimate goal of research; after all, researchers are really interested in populations, not samples. Table 8.1 lists the three types of statistical procedures and their respective main purposes.

Reading Question

4. The purpose of a confidence interval is to
   a. test for significant differences.
   b. quantify the effectiveness of a treatment.
   c. help researchers estimate a population parameter with a specific level of confidence and precision.

Table 8.1 Statistical Procedures and Their Main Purposes

Statistical Procedure	*Main Purpose*
Significance/hypothesis testing	To assess the probability of a result being created by sampling error; if the null is rejected, it is probably not sampling error
Effect size	To quantify a treatment's effectiveness
Confidence intervals	To help estimate a population parameter with a specific level of confidence and precision

Reading Question

5. Confidence intervals provide researchers with
   a. a range of plausible values for a population parameter.
   b. information that helps apply results from samples to populations.
   c. both of the above.

LOGIC OF CONFIDENCE INTERVALS

As stated earlier, the purpose of confidence intervals (CIs) is to help researchers apply sample results to populations. More specifically, CIs provide a range of plausible values for a population parameter. This range of plausible values is defined by the upper and lower boundary values of the CI. The **upper boundary** *is the largest plausible parameter value,* and the **lower boundary** *is the smallest plausible parameter value*. CIs use statistics from samples to determine these boundary values. All values between these boundary values, including the boundary values themselves, are considered plausible parameter values for the population.

While there are many different kinds of CIs, all of them have a similar logic. Every CI uses a point estimate and a margin of error around that point estimate to find upper and lower boundary values. When estimating a population mean, the **point estimate** *is the sample mean, which is the single most plausible value for* μ, and a **margin of error** *determines the range of additional plausible values for that parameter*. In the above example, the sample mean for college students' weekly exercise time after taking the healthy living class ($M = 155.67$) is the most plausible prediction for the mean exercise time of the population if they take the healthy living class. But, because of sampling error, this point estimate of 155.67 is probably not perfectly accurate. Therefore, a margin of error is added to and subtracted from the point estimate to obtain the upper and lower boundaries of the CI. The conceptual formulas for the upper and lower boundaries are given as follows:

$$\text{Upper boundary} = \text{Point estimate} + \text{Margin of error}.$$

$$\text{Lower boundary} = \text{Point estimate} - \text{Margin of error}.$$

Reading Question

6. Which of the following might be used as a point estimate in a CI?
   a. Upper bound
   b. Lower bound
   c. Sample mean
   d. Margin of error

Reading Question

7. The size of the range of plausible parameter values provided by a CI is determined by the
   a. sample mean.
   b. margin of error.

The center of a CI is the sample mean. The size of the margin of error around the sample mean is determined by two factors: (1) the expected amount of sampling error (*SEM*) and (2) the specific level of confidence you want to achieve (usually either 95% or 99% confidence). It probably makes sense to you that studies with larger amounts of sampling error have CIs with larger margins of error. After all, with more sampling error, more values are plausible for the population mean. It might be less obvious that if researchers want more confidence in their estimate, their margin of error must be larger. By way of analogy, suppose your instructor asks you to estimate your score on the upcoming statistics final. You could say, "I'm going to get a score between 84% and 86%," or you could say, "I'm going to get a score between 70% and 100%." The first estimate is more precise (i.e., has fewer plausible values). Precise estimates have a small range of plausible values (i.e., margin of error); therefore, they have to be made with very low confidence. Conversely, the second estimate, a grade between 70% and

100%, is less precise but it can be made with greater confidence. There is always an inverse relationship between the precision of an estimate (i.e., how narrow the range of values are) and the confidence one should have in that estimate. Therefore, when all other things are equal, if you want to be 99% confident in your estimate, your CI will need to be *wider* than if you only need 95% confidence. More confidence requires a wider margin of error.

Reading Question

8. Which of the following contribute to the size (i.e., width) of the margin of error? (Choose two.)
   a. Sample mean
   b. Confidence level (95% vs. 99%)
   c. Expected sampling error

Reading Question

9. In general, 99% CIs are _____ than 95% CIs.
   a. wider
   b. narrower
   c. more precise

COMPUTING A CONFIDENCE INTERVAL FOR A POPULATION MEAN

The formulas in the previous section are "conceptual" formulas in that they help explain the general logic of CIs. The formulas described in this section will be the ones you actually use when estimating a population's mean. As mentioned earlier, a sample mean serves as the point estimate, and a margin of error is added to and subtracted from the sample mean to find a CI's upper and lower boundaries. In this section, you will learn to compute the margin of error. As mentioned above, the margin of error is derived from the expected sampling error, *SEM*, and a specific level of confidence (usually 95% or 99%).

When using a sample mean to estimate a population's mean, the expected amount of sampling error is defined by the SEM_s $\left(\text{i.e., } \frac{SD}{\sqrt{N}}\right)$. The specific level of confidence (95% or 99%) is defined by a specific t value from the critical t table. As always, the specific critical t value depends on the degrees of freedom (df) associated with the sampling error formula, in this case, $df = N - 1$. As indicated in Table 8.2, when computing a 95% CI, use the df to look up the critical t value in the two-tailed .05 critical t table. A 99% CI requires the critical t value in the two-tailed .01 critical t table. The activity in this chapter will explain why these specific values are used for 95% and 99% CIs, respectively. At this point, it is enough that you know when to use which t table.

Table 8.2 Critical t Values for Desired CI

Desired CI	*Critical t Value to Obtain From Table*
95% CI	Use two-tailed .05 critical t
99% CI	Use two-tailed .01 critical t

Reading Question

10. When creating a 95% CI, use a critical t value from the
    a. two-tailed .01 t table.
    b. two-tailed .05 t table.
    c. one-tailed .01 t table.
    d. one-tailed .05 t table.

In the example at the beginning of this chapter, we used a single-sample *t* test to compare the mean weekly exercise time of college students who took a healthy living class, $M = 155.67$, to 150 minutes. We found that those who took the class exercised more ($M = 155.67$, $SD = 20.85$) than the population mean of 150 minutes, $t(59) = 2.11$, $p = .04$, $d = .27$. Now, the provost wants to apply these research findings to the population of interest to him—namely, his college students. He wants to know what the range of plausible values for the mean weekly exercise time would be *if the population took the healthy living class*. He needs you to compute a 95% CI to estimate this population parameter. Start by writing down the following formulas for the upper and lower boundaries of the CI.

$$\text{Upper boundary} = M + (t_{CI})(SEM_s).$$

$$\text{Lower boundary} = M - (t_{CI})(SEM_s).$$

For both equations, you need the same three values: (1) the point estimate or *M*, (2) the correct critical *t* value for the CI or t_{CI}, and (3) the expected sampling error or SEM_s. The point estimate comes from the sample mean ($M = 155.67$). The critical *t* score comes from the two-tailed .05 critical *t* table because we are computing a 95% CI. The *df* is 59 ($df = n - 1 = 60 - 1 = 59$), so the t_{CI} is 2.001. And the SEM_s is $\frac{SD}{\sqrt{N}}$ or $\frac{20.85}{\sqrt{60}} = 2.69$. Therefore, the upper boundary of the 95% CI is

$$\text{Upper boundary} = M + (t_{CI})(SEM_s).$$

$$\text{Upper boundary} = 155.67 + (2.001)(2.69).$$

$$\text{Upper boundary} = 155.67 + (5.383) = 161.05.$$

You should notice that the final line of the above calculation is identical to the conceptual formula discussed in the previous section, the point estimate of 155.67 plus the margin of error of 5.383. The lower boundary is

$$\text{Lower boundary} = M - (t_{CI})(SEM_s).$$

$$\text{Lower boundary} = 155.67 - (2.001)(2.69).$$

$$\text{Lower boundary} = 155.67 - (5.383) = 150.29.$$

Again, it is worth noticing that the lower boundary is the point estimate of 155.67 minus the margin of error of 5.38. Based on these computations, we should be 95% confident that the actual mean exercise time of the population of college students if they took the healthy living class would be between 150.29 and 161.05 minutes per week. This CI might help policy makers determine if the health benefits gained by requiring the course would be worth the added cost of offering the course. CIs help researchers, or policy makers, think about the possible consequences of applying results from a sample to an entire population.

Reading Question

11. Which of the following determine the correct t_{CI} score for a given CI? (Choose two.)
    a. Confidence level (e.g., 95% or 99%)
    b. The *df*
    c. Expected sampling error

Reading Question

12. In the above example, the upper and lower CI boundaries represent
    a. the highest and lowest values for the population parameter that are considered plausible, or reasonable considering sampling error.
    b. the effect size from the study.
    c. the probability of rejecting the null hypothesis.

COMPUTING CONFIDENCE INTERVALS FOR A MEAN DIFFERENCE

In the previous section, we computed a CI that provided a range of plausible values for a *population mean*. We also learned that CIs provide very useful information on how sample results might generalize to an entire population. In this section, we will compute a different kind of CI that provides different but equally useful information to researchers and policy makers. This CI estimates the *mean difference* rather than the mean itself. For example, suppose policy makers wanted to know how much weekly exercise time would *change* in the population if the healthy living class was required.

Like all other CIs, this one is computed with a point estimate, an expected sampling error, and a t_{CI}. The point estimate, in this case, is the *mean difference* between the sample mean ($M = 155.67$) and the population mean ($\mu = 150$), 5.67 (i.e., 155.67 − 150.00 = 5.67). The expected sampling error (SEM_s) is the same as in the previous example. The t_{CI} value is also the same. So, the only difference between this CI and the one computed above is that we are using a difference between a sample mean and a population mean as the point estimate rather than the sample mean itself. The upper and lower boundaries are worked out as follows:

$$\text{Upper boundary} = (\textit{point estimate}) + \text{Margin of error}.$$

$$\text{Upper boundary} = (M - \mu) + (t_{CI})(SEM_s).$$

$$\text{Upper boundary} = (155.67 - 150) + (2.001)(2.69).$$

$$\text{Upper boundary} = 5.67 + (5.383) = 11.05.$$

$$\text{Lower boundary} = (\textit{point estimate}) - \text{Margin of error}.$$

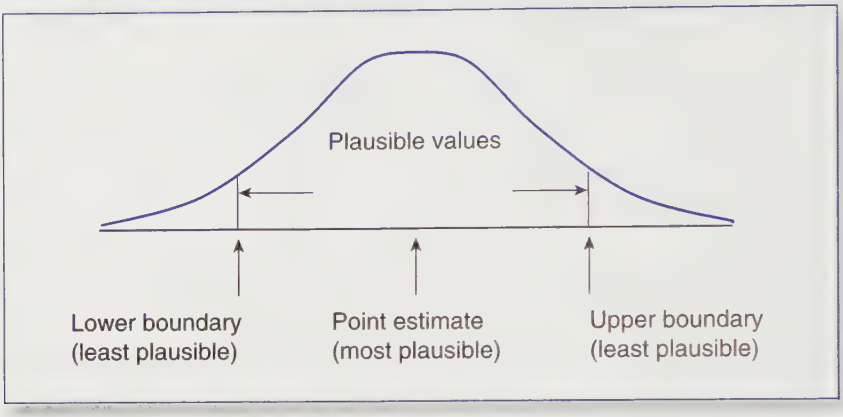
$$\text{Lower boundary} = (M - \mu) - (t_{CI})(SEM_s).$$

$$\text{Lower boundary} = (155.67 - 150) - (2.001)(2.69).$$

$$\text{Lower boundary} = 5.67 - (5.383) = .29.$$

Based on these computations, we should be 95% confident that if the population of college students took the healthy living class, their exercise time would increase by between .29 and 11.05 minutes each week. Now, how should you interpret this range of plausible values? First, you should recognize that not all of the CI values are equally plausible. Figure 8.1 illustrates the relative plausibility of each value in the CI. While all of the values in the CI are considered plausible, the upper and lower boundary values are the least plausible and the point estimate is the most plausible. The general rule is that CI values become less plausible with increased distance from the point estimate.

Figure 8.1 Plausibility of CI Values



Second, whether or not the mean difference of zero is included in the range of plausible values is worth noticing. If a mean difference of zero is not between the upper and lower boundaries of the 95% CI for the mean difference, the sample mean of 155.67 and the population mean of 150 are significantly different from each other. This conclusion is completely consistent with the results of the significance test that was described at the beginning of the chapter. This will always be the case. If you conduct a two-tailed significance test with an alpha value of .05 and you reject the null, the 95% CI for the mean difference will never contain zero. In contrast, if you fail to reject the null, the 95% CI will always contain zero.

Reading Question

13. When computing a CI for a mean difference, if zero is found to be a plausible value (i.e., it is between the upper and lower boundaries), a two-tailed significance test would find that the two means are
    a. not significantly different.
    b. significantly different.

Why should you conduct a CI for a mean difference after conducting a significance test? The answer harkens back to the beginning of the chapter. The two statistical procedures have distinct purposes. The significance test indicates whether or not a given effect was likely created by sampling error. Then, the effect size describes the effect's magnitude. And finally, the CI helps you generalize your results to the population by identifying plausible population parameter values.

Reading Question

14. Which of the following determines plausible values for population parameters?
    a. Significance testing
    b. Effect sizes
    c. CIs

This chapter illustrates how to compute two different types of CIs. Table 8.3 displays the two types of CIs, the appropriate *df* formula, and the specific CI formula.

Table 8.3 CI Types, Their Degrees of Freedom, and Specific Formulas

CI Type	df *Formula*	*Specific CI Formula*
Estimating μ	$df = N - 1$	$UB = M + (t_{CI})\left(\frac{SD}{\sqrt{N}}\right)$ $LB = M - (t_{CI})\left(\frac{SD}{\sqrt{N}}\right)$
Estimating mean difference of sample mean (M) and population mean (μ)	$df = N - 1$	$UB = (M - \mu) + (t_{CI})\left(\frac{SD}{\sqrt{N}}\right)$ $LB = (M - \mu) - (t_{CI})\left(\frac{SD}{\sqrt{N}}\right)$

Note: CI = confidence interval; *df* = degrees of freedom; UB = upper boundary; LB = lower boundary.

REPORTING CONFIDENCE INTERVALS IN APA STYLE

While the CI procedure is not new, its prominence within the behavioral sciences has been increasing. Evidence for the increasing importance of CIs is found in the most recent version of the American Psychological Association (APA) publication manual. The manual strongly recommends that researchers compute and interpret CIs in conjunction with significance tests and effect sizes when reporting results (APA, 2010, p. 34). Specifically, the APA publication manual recommends that authors use CIs when discussing the implications of their research findings (i.e., when applying their results to populations).

The following format is consistent with the latest APA manual recommendations for reporting significance tests, effect sizes, and CIs. Note how the CI information we computed in this chapter is integrated into the report. It reports the 95% CI for the population mean and the mean difference. The first time you report a CI for a set of analyses, you report the level of confidence (i.e., 95% or 99%). After that, you do not need to indicate the level of confidence because the reader can infer that you used the same level of confidence for all analyses.

A single-sample *t* test revealed that students who took the healthy living class exercised more weekly ($M = 155.67$, $SD = 2.69$), 95% CI [150.29, 161.05], than students in the population ($\mu = 150$), $t(59) = 2.11$, $p = .04$, $d = .27$, CI [0.29, 11.05].

Reading Question

15. The APA manual recommends that researchers use
    a. CIs rather than significance testing.
    b. CIs in conjunction with significance testing and effect sizes.

CONFIDENCE INTERVALS FOR EFFECT SIZES

The APA also recommends that CIs be reported for every effect size. Unfortunately, these computations are complex and beyond the scope of this book. Cumming (2012) provides detailed, yet extremely

readable, discussions of CIs for effect sizes. He also provides a free Microsoft Excel spreadsheet for computing and graphing many types of CIs, including those for effect sizes. We highly recommend Cumming's book to readers wanting a greater understanding of CIs.

INTERPRETATIONS OF CONFIDENCE INTERVALS

The main purpose of CIs is to help researchers apply results from a sample to a population. Cumming (2012) identifies six ways this could be done. We will present four ways to interpret CIs.

First, a 95% CI suggests that you can be 95% confident that the true population mean falls between the lower and upper boundaries, including the boundary values themselves. In other words, a CI can be interpreted as *a set of plausible values for the population parameter* (i.e., a population mean or a population mean difference, depending on the type of CI). The point estimate is considered the most plausible value, and the plausibility of each value decreases as you move away from the point estimate. Values *outside* of the upper and lower boundaries are interpreted as *implausible*.

Second, the width of CIs helps researchers interpret the precision of a given study's parameter estimate. *Narrower CIs (i.e., those with smaller margins of error) provide more precise estimates.*

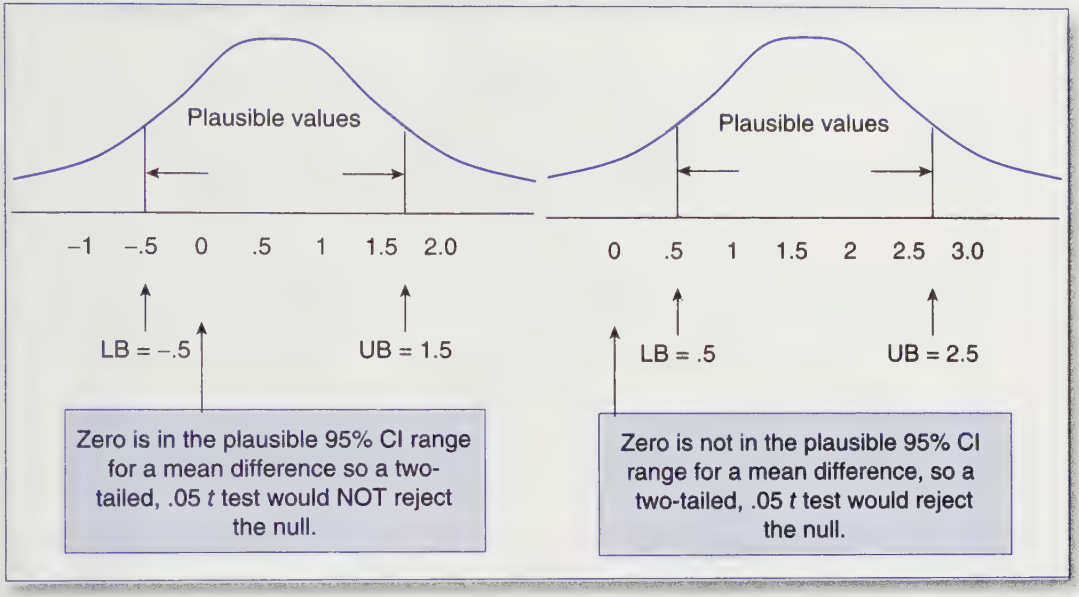
Third, *a 95% CI has a .83 replication recapture rate*. This interpretation requires some more explanation. If a given study produces a 95% CI of [10, 20], the first CI interpretation described earlier indicates that researchers should be 95% confident that μ is between 10 and 20, inclusively. However, this does not mean that if the researchers repeated the study a second time, the mean of the second study would be between 10 and 20, 95% of the time. In fact, assuming that the second study was done exactly as the first, the mean of the second study is expected to be between 10 and 20 only 83% of the time. This is referred to as the **replication recapture rate,** *the probability that the point estimate of a study that is replicating a previous study will fall between the upper and lower boundaries of the previous study*. So the third interpretation is that if a given study is replicated, 83% of the time the point estimate of the second study will fall between the boundaries of the original study's 95% CI.

Finally, we'll mention an additional way to use CI logic that is limited to CIs for *mean differences*. It is a logical extension of the first interpretation we discussed previously—namely, that values *outside* the CI boundaries are *implausible*. When computing a 95% CI for a *mean difference*, if zero is not plausible, a two-tailed .05 *t* test would find that the two means are significantly different. On the other hand, if zero is located between the upper and lower boundaries, the two means would not be significantly different. For example, consider a 95% confidence interval for a mean difference with a lower bound of −.5 and an upper bound of 1.50. In Figure 8.2a, the point estimate is the sample mean difference of .5. The lower bound is −.5 and the upper bound is 1.5. A mean difference of zero falls between the upper and lower bound. For this CI, a mean difference of zero is a plausible value, so if you ran a two-tailed *t* test with $\alpha = .05$ to determine if this difference was significant, you would not reject the null hypothesis. However, as illustrated in Figure 8.2b, a confidence interval with a point estimate of 1.5, a lower bound of .5, and an upper bound of 2.5 does not include zero, and so if you ran a two-tailed test with $\alpha = .05$ on that data, you would reject the null hypothesis. Figures 8.2a and 8.2b should help you understand the relationship between the results of a CI and the results for significance test for a mean difference. If zero is between the lower and upper boundaries, the difference is not significant. If it is outside these boundaries, it is significant.

Reading Question

16. Which of the following is not a correct interpretation of a 95% CI?
    a. A range of plausible parameter values
    b. A wider CI is less precise than a narrow CI
    c. A 95% CI of [15, 30] means that 95% of the replications of that study will produce a sample mean between 15 and 30.

Figure 8.2 Example of Zero Mean Difference Being (a) Plausible and (b) Implausible



Reading Question

17. If the value of zero is not included between the upper and lower boundaries of a 95% CI for a *mean difference,* the two means used to create the point estimate are
    a. not significantly different.
    b. significantly different.

SPSS

CI for a Mean

SPSS will compute a 95% or 99% CI for a mean through the Explore menu.

- Click on Analyze, Descriptive Statistics, and then Explore.
- Move the variable of interest (in this case, it is ExerciseMinutes) into the Dependent List box.
- Click on the Statistics button.
- In the Explore:Statistics box, select Descriptives and make sure the Confidence Interval for Mean is set at 95%.
- Click on the Continue button.
- Under Display, select Statistics.
- Click on the OK button to run the analysis.

Figure 8.3 displays the SPSS output produced by the above steps.

Figure 8.3 Annotated SPSS Output for a 95% Confidence Interval for a Mean

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
ExerciseMinutes	60	100.0%	0	.0%	60	100.0%

Descriptives

			Statistic	Std. Error
ExerciseMinutes	Mean		155.6667	2.69134
	95% Confidence Interval for Mean	Lower Bound	150.2813	
		Upper Bound	161.0520	
	5% Trimmed Mean		155.3148	
	Median		153.5000	
	Variance		434.599	
	Std. Deviation		20.84704	
	Minimum		109.00	
	Maximum		205.00	
	Range		96.00	
	Interquartile Range		30.00	
	Skewness		.263	.309
	Kurtosis		-.335	.608

CI for a Single-Sample Mean Difference

SPSS automatically computes 95% confidence intervals for *mean differences* every time you run a single-sample *t* test. If you want to compute the 99% confidence interval around the mean difference, click Options and change the confidence interval to 99%. Figure 8.4 displays the SPSS output for this analysis.

Reading Question

18. The above output displays which of the following?
    a. A 95% CI for a sample mean
    b. A 95% CI for two sample means
    c. A 95% CI for the mean difference of samples

OVERVIEW OF THE ACTIVITY

In Activity 8.1, you will compute confidence intervals for means and mean differences using a calculator and SPSS. This activity also highlights the differences among significance testing, effect sizes, and confidence intervals.

Figure 8.4 Annotated SPSS Output for a 95% Confidence Interval for a Mean Difference

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
ExerciseMinutes	60	155.6667	20.84704	2.69134

One-Sample Test

	Test Value = 150					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
ExerciseMinutes	2.106	59	.040	5.66667	.2813	11.0520

Lower boundary of 95% CI for mean difference

Activity 8.1: Estimating Sample Means and Sample Mean Differences

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Summarize the key points of significance testing
- Construct 95% and 99% confidence intervals
- Interpret confidence intervals correctly
- Compute confidence intervals using SPSS
- Correctly interpret confidence intervals
- Include confidence intervals in APA style write-ups
- Correctly interpret the statistical information in an APA style write-up

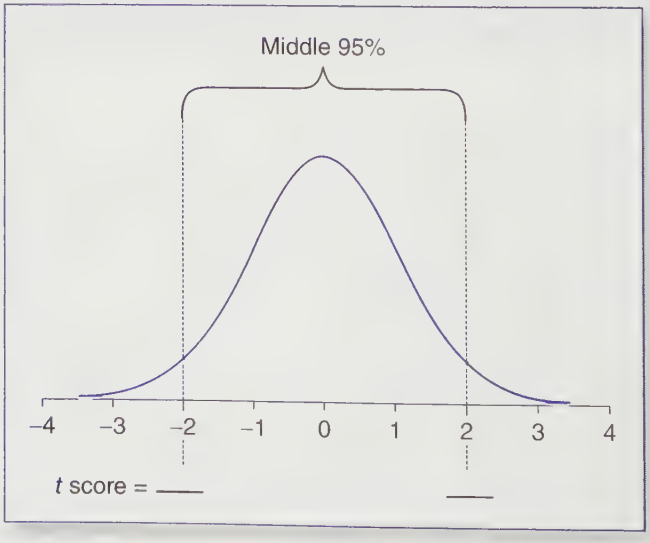
INTRODUCTION TO CONFIDENCE INTERVALS

As you know from the chapter, a confidence interval is a different statistical procedure that is used for a different purpose than hypothesis testing or effect sizes. Its main purpose is estimating the value of a population parameter. For example, suppose you just took a job as an English teacher in a large school district. Knowing the mean reading comprehension score of all seniors in your school district would help you understand the current state of education in your district as well as help you set achievement goals for your students. In this situation, you could use a sample of 50 seniors' standardized reading comprehension scores to estimate the mean reading comprehension score for the entire population. Of course, we have used samples to represent populations throughout this book. However, how much confidence should you have that the sample mean is a good estimate of the actual population parameter? Confidence intervals answer this question. A **confidence interval** is *a range of values that contain the actual population parameter with a specific degree of certainty or confidence*. For example, in this situation, you could compute a 95% confidence interval of 29 to 43. *This confidence interval would mean that you could be 95% confident that the actual mean reading comprehension score for the population of seniors in your district is between 29 and 43.*

In this activity, you will compute and interpret 95% and 99% confidence intervals (CIs) for individual means and for mean differences. However, there is some preliminary information you should understand before you can start computing confidence intervals.

Review of *t* Distributions

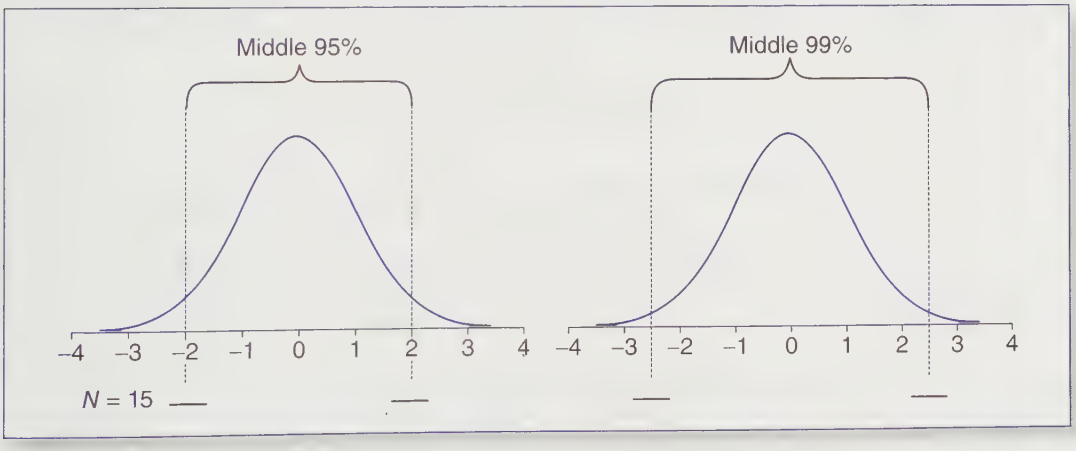
1. As you know, scores in the middle of *t* distributions are more common than scores in the tails. The following *t* curve represents all possible sample mean differences if the null hypothesis is true. The two vertical lines create three areas on the curve; if the middle section contains the middle 95% of all *t* scores (and all sample mean differences), what percentage of all *t* scores are in the left and right tails of the distribution?
   a. 5% in the left and 5% in the right
   b. 2.5% in the left and 2.5% in the right
   c. 1% in the left and 1% in the right



2. To compute a 95% confidence interval, you must know the two *t* scores that bracket the middle 95% of all possible *t* scores. In the previous question, you determined that the left and right tails contained the bottom and top 2.5% of all *t* scores, respectively. Now we need to find the *t* scores that define the bottom and top 2.5% of *t* scores. Which table of critical *t* values do you use to find the negative and positive *t* scores that define the bottom and top 2.5% of *t* scores?
   a. One-tailed $\alpha = .05$ table
   b. Two-tailed $\alpha = .05$ table
   c. One-tailed $\alpha = .01$ table
   d. Two-tailed $\alpha = .01$ table

3. As you know, the *t* distribution changes shape as sample size changes. Therefore, if you want to know the *t* scores that define the middle 95% of all possible *t* scores (and all possible sample means), you need to know the sample size so you can compute the degrees of freedom and then find the correct *t* score. If the sample size is 50 and $df = N - 1$, what are the *t* scores that define the middle 95% of sample means? Find these values and place them on the figure on page 254.

4. Complete the following figures and charts by first determining which critical *t* table you need to use to find the middle 95% of sample means and middle 99% of sample means. After you determine which table to use for each middle percentage, use the provided sample size to find the specific *t* values for each middle percentage.



Sample Size	*Middle % of* t *Scores/Sample Means*	*Critical* t *Table to Use*	*Specific* t *Scores Defining the Middle %*
$N = 15$	95%	Two-tailed .05	
$N = 15$	99%		

5. Now complete the following table for a larger sample size.

Sample Size	*Middle % of* t *Scores/Sample Means*	*Critical* t *Table to Use*	*Specific* t *Scores Defining the Middle %*
$N = 50$	95%		
$N = 50$	99%		

STATISTICAL ESTIMATION

As mentioned above, in some situations, the goal is estimating the value of a population parameter rather than testing a hypothesis. In the hypothetical example outlined earlier, you are an English teacher in a large school district and you want to know the mean reading comprehension score of all seniors in your district.

6. Let's assume that the entire population of seniors is too large for you to actually measure every senior's reading comprehension ability. What could you do to estimate the mean reading comprehension score of seniors from this district?
   a. Nothing. You can only estimate the sample statistic when you are working with a sample.
   b. Compute the effect size and determine if it is small, medium, or large. The larger the effect size, the more accurately it estimates the population parameter.
   c. Obtain a sample and compute the mean reading comprehension score from the sample. Then, compute the confidence interval around the sample mean.

7. Fill in the following blanks with the following terms: confidence interval, sampling error, and population mean.

   You decide to collect a random sample of 50 seniors and give them a standardized reading comprehension test. Based on the central limit theorem, the mean of this sample is expected to be equal to the _____ _____. However, because of _____ _____ you would not expect the mean of the sample to be exactly equal to this value. In this situation, you can use the expected amount of sampling error to construct a range of values that probably contains the actual mean reading comprehension score of the population (i.e., the population parameter). This range of values is called a _____ _____.

Confidence intervals have three components: (a) a point estimate, (b) a specific confidence level, and (c) an expected amount of sampling error. Point estimates come from samples. In this case, the sample mean estimates the population mean; the sample mean is the point estimate. Confidence levels come from the *t* scores that define the middle of the *t* distributions (i.e., like those values you found in the above tables). *If you want to be 95% confident that your interval of values contains the true population mean, you would use the* t *scores that contain the middle 95% of sample means.* If you want

to be 99% confident, you would use the t scores that correspond to the middle 99%. Finally, the expected sampling error is computed using the *SEM* formula you have learned. In this situation, we are estimating the *value of a population mean,* so you would use the *SEM* from the single-sample t test $\left(\text{i.e.,}\frac{SD}{\sqrt{N}}\right)$. The specific formulas for constructing the upper and lower boundaries of a confidence interval for a population mean are

$$\text{Upper boundary} = M + (t_{CI})(SEM_s).$$

$$\text{Lower boundary} = M - (t_{CI})(SEM_s).$$

Therefore, if $N = 50$, $M = 36$, and $SD = 5$, the upper boundary of a 95% confidence interval would be

$$\text{Upper boundary} = 36 + (2.0096)\left(\frac{5}{\sqrt{50}}\right) = 37.42.$$

8. Compute the lower boundary.

$36 - (2.0096)\left(\frac{5}{\sqrt{50}}\right) = 34.57$

9. The proper interpretation of this confidence interval is the following:
   a. We are 95% confident that the actual mean reading comprehension score for the entire population of seniors in the school district is between the lower and upper boundary values of 34.57 and 37.42.
   b. We are 95% confident that the true population mean is 36.

COMPUTING CONFIDENCE INTERVALS: BASIC PROBLEMS

For the following problems, use the information provided to compute the appropriate confidence interval.

Scenario 1

$M = 59$; $SD = 4$; $N = 15$; compute a 95% confidence interval; you are estimating a population mean, so use $\frac{SD}{\sqrt{N}}$ for sampling error. 1.03

10. What is the point estimate?

11. Compute the margin of error.

$$\frac{4}{\sqrt{15}}$$

12. What is the upper boundary for the confidence interval?

13. What is the lower boundary for this confidence interval?

14. Choose the correct interpretation of this confidence interval.
    a. We are confident that 95% of the scores are between 56.79 and 61.22.
    b. We are 95% confident that the sample mean is between 56.79 and 61.22.
    c. We are 95% confident that the population mean is between 56.79 and 61.22.

Scenario 2

$M = 59$; $SD = 4$; $N = 15$; compute a 99% confidence interval; you are estimating a population mean, so use $\frac{SD}{\sqrt{N}}$ for sampling error.

15. What is the point estimate?

16. Compute the margin of error.

17. What is the upper boundary for this confidence interval?

18. What is the lower boundary for this confidence interval?

19. Choose the correct interpretation of this confidence interval.
    a. If this study was replicated 100 times, 99% of the means would fall between 55.93 and 62.07.
    b. We are 99% confident that the population mean increased by between 55.93 and 62.07.
    c. We are 99% confident that the population mean is between 55.93 and 62.07.

Scenario 3

$M = 59$; $SD = 2$; $N = 15$; compute a 95% confidence interval; you are estimating a population mean, so use $\frac{SD}{\sqrt{N}}$ for sampling error.

20. What is the point estimate?

21. Compute the margin of error.

22. What is the upper boundary for this confidence interval?

23. What is the lower boundary for this confidence interval?

24. Choose the correct interpretation of this confidence interval.
    a. We can be 95% confident that the population mean is between the upper and lower bounds of the CI.
    b. If we were to take 100 additional samples from the same population, 95% of the confidence intervals would contain the sample mean.

25. When all other factors are held constant, as sampling error gets smaller, what happens to the width of confidence intervals?
    a. They get wider.
    b. They get narrower.
    c. It stays the same.

26. When all other factors are held constant, as confidence levels decrease (i.e., 99% to 95%), what happens to the width of confidence intervals?

a. It gets wider.
b. It gets narrower.
c. It stays the same.

27. Not surprisingly, there is a trade-off when constructing confidence intervals. Generally speaking, higher confidence levels are associated with less precision (i.e., wider confidence intervals). If you require a high degree of confidence in your estimate, the interval will be _____ (choose one: wider, narrower). In contrast, if you desire a high degree of precision in your estimate, your confidence level will be _____ (choose one: lower, higher).

CONFIDENCE INTERVALS FOR MEAN DIFFERENCES

Scenario 4

All the confidence intervals you constructed thus far were estimating a population mean. You can also compute a confidence interval around a mean difference. For example, in a test of scientific literacy, students are asked to indicate how long *Homo sapiens* have been on Earth. Based on current scientific evidence, the best answer to that question is 200,000 years. The average answer given by a sample of 50 students was 250,000 with a standard deviation of 500,000 years. Compute a 95% confidence interval.

28. In this case, the point estimate is the difference between the sample mean and 200,000. What is your point estimate for the *mean difference* in this scenario?

29. What is the margin of error in this scenario?

30. What is the upper boundary for this confidence interval?

31. What is the lower boundary for this confidence interval?

32. Choose the correct interpretations of this confidence interval. Select all that apply.
    a. 95% of the sample mean differences are between –92,100.18 and 192,100.18.
    b. If multiple samples were taken from this same population, 95% of the confidence intervals would contain the population mean difference.
    c. We are 95% confident that the true value of the population mean difference is between –92,100.18 and 192,100.18.
    d. 95% of the respondents gave answers that were between –92,100.18 and 192,100.18.

CONFIDENCE INTERVALS IN SPSS

In the previous questions, you learned to compute confidence intervals around individual means and around mean differences. Although there may be times when you need to compute confidence intervals by hand, you can also obtain confidence intervals using SPSS. The following examples will show you how to use SPSS to compute confidence intervals and how to incorporate confidence intervals into an APA-style write-up.

Estimating a Population's Parameter

A huge international corporation, IM-ANX, has hired your consulting firm to conduct a study on its employees' job satisfaction. While the corporation wants all of its employees to be satisfied, it is particularly interested in keeping its highly trained employees satisfied. These highly trained employees (e.g., biochemical engineers, genetic researchers) are very hard to find, and it seriously hurts the corporation when these employees quit. The corporation hired your firm to determine how satisfied these highly trained employees are currently and to suggest ways that might increase these employees' satisfaction. As a first step, your firm obtains a random sample of 35 "highly trained" employees from the corporation. Each of these 35 employees completes a job satisfaction survey that measures the degree to which they agree with 20 different statements about their job satisfaction using a 5-point Likert response scale. Two examples of the 20 questions are provided as follows:

1. I am satisfied with my salary.

1	2	3	4	5
Strongly disagree		Neither agree nor disagree		Strongly agree

2. Overall, I am satisfied with my job.

1	2	3	4	5
Strongly disagree		Neither agree nor disagree		Strongly agree

33. After collecting the 35 employees' responses to all 20 questions, you create a single overall job satisfaction variable by *averaging* each employee's responses to all 20 questions. So each employee now has a single score reflecting his or her overall job satisfaction. What is the lowest possible score and the highest possible score on this overall job satisfaction variable?

Lowest = 1 (meaning not at all satisfied)

Highest = 5 (meaning very satisfied)

34. The first question your firm needs to answer is, "How satisfied is the *population* of highly trained employees currently?" Obviously, you only have data from the sample of 35 employees. Which of the following statistical procedures will help you estimate the *population's* level of job satisfaction (i.e., the population parameter)?

a. A significance test
b. An effect size
c. A confidence interval

35. Some interns at your firm entered the data into an SPSS file titled "JobSatisfaction.sav." Obtain this data file and then compute the mean level of job satisfaction for the sample of 35 employees and the 95% confidence interval around this mean. You could do this by hand, but you can also get confidence intervals through SPSS. To obtain the mean and the 95% confidence interval using SPSS:

- Click on Analyze, Descriptive Statistics, and then Explore.
- Move the variable of interest (in this case, it is JobSatisfaction) into the Dependent List box.
- Click on the Statistics button.
- In the Explore:Statistics box, select Descriptives and make sure the Confidence Interval for Mean is set at 95%.
- Click on the Continue button and then on the OK button to run the analysis.

What is the mean level of job satisfaction in the sample of 35 employees?

36. The sample mean indicates that the sample had a mean job satisfaction score of 3.4. This value suggests the sample is slightly satisfied with their jobs; 3.4 is slightly higher than the neutral point of 3 on the scale. But, this is just the sample. This sample might not represent the population very well. The actual population's mean might be higher or lower due to _____ _____. To account for this possibility, your firm should compute *the range of plausible values* for the *population's* mean job satisfaction. This is exactly what the 95% confidence interval gives you.

37. Find the lower and upper bounds for the 95% CI for the population's mean job satisfaction in the SPSS output.

a. LB = −.8412, UB = .2412
b. LB = 2.8588; UB = 3.9412
c. LB = 1.00; UB = 5.00
d. LB = 1.5755; UB = 3.4000

38. Which of the following is a correct interpretation of this confidence interval?

a. We are 95% confident that the true value for the population mean level of job satisfaction is between 2.86 and 3.94.
b. We are 95% confident that the true value for the population mean difference between the employee's level of job satisfaction and the midpoint of the scale (3) is between 2.86 and 3.94.
c. There is a 95% chance that the confidence interval contains the sample mean.
d. There is a 95% chance that the sample mean is between the lower and upper boundary values.

39. Job satisfaction was measured on a 5-point scale, with 3 meaning moderate satisfaction. Based on the 95% CI, the *population* of highly trained employees at IM-ANX are
    a. likely to be so satisfied there is really no way to make them more satisfied.
    b. likely to be slightly satisfied, but there is certainly room to increase their satisfaction.
    c. likely to be unsatisfied with their jobs; they might be on the verge of quitting.

40. The lower bound of the 95% CI suggests that
    a. it is plausible that the population's mean satisfaction score is below the midpoint on the scale.
    b. it is not plausible that the population's mean satisfaction score is below the midpoint on the scale.

Comparing Means

41. You've learned that IM-ANX's highly skilled employees are likely to be "slightly satisfied." While this is useful information, you know they will want to know more. For example, "how satisfied are their employees compared to a competitor's employees?" By reviewing some professional research articles on job satisfaction, you learn that the mean job satisfaction ratings for all "highly trained" employees at large corporations is 3.70 on the 20-question job satisfaction survey. Obviously, this 3.70 is higher than the *estimate* of IM-ANX's mean satisfaction of 3.40, but this small difference might be due to sampling error. Which of the following statistical procedures should you use to determine if this difference between 3.70 and 3.40 is likely or unlikely to be due to sampling error?
    a. Significance testing
    b. Effect size
    c. Confidence interval

42. Which of the following significance tests should you use to determine if the difference between all corporations' mean of 3.70 (a population parameter) and the sample mean of 3.40 is likely or unlikely to be due to sampling error?
    a. Single-sample *t* test
    b. *z* for a sample mean
    c. *z* for a single score

43. Next, you will need to compute a single-sample *t* significance test to determine if the sample mean of 3.40 is significantly different from the large corporation mean of 3.7. Refer to Chapter 7 for instructions on how to use SPSS to run the single-sample *t* test. What *t* value did you obtain from SPSS?
    a. 3.4000
    b. −1.126
    c. −.3000
    d. .268

44. You plan to compare the average level of job satisfaction at IM-ANX to 3.7 (the mean for all large corporations). In this case, a one-tailed test is more appropriate because you only want to know if job satisfaction is significantly lower than 3.7. Write H_0 next to the null hypothesis and H_1 next to the research hypothesis for this *t* test.

a. _____IM-ANX's average job satisfaction score for all employees is not lower than 3.7 (the large corporation mean).

b. _____IM-ANX's average job satisfaction score for all employees is lower than 3.7 (the large corporation mean).

c. _____IM-ANX's average job satisfaction score for all employees is equal to 3.7 (the large corporation mean).

d. ____IM-ANX's average job satisfaction score for all employees is not equal to 3.7 (the large corporation mean).

45. What is the p value for this one-tailed significance test?

a. .268

b. .134

c. .05

46. Choose the statement that is a correct interpretation of this p value.

a. The probability that the research hypothesis is true is .134.

b. The probability that the null hypothesis is true is .134.

c. The probability that the researcher made a Type I error is .134.

d. If the null hypothesis is true, the probability of obtaining a t value of −1.126 or lower is .134.

47. Should you reject the null hypothesis (use $\alpha = .05$)?

a. Yes, $p < .05$

b. No, $p > .05$

48. Which of the following is the best verbal summary of the results of the significance test?

a. The mean job satisfaction of the sample of highly trained employees at IM-ANX is not significantly different from the overall corporation satisfaction mean of 3.7.

b. The mean job satisfaction of the sample of highly trained employees at IM-ANX is significantly different from the overall corporation satisfaction mean of 3.7.

c. The mean job satisfaction of the sample of highly trained employees at IM-ANX is not significantly lower than the overall corporation satisfaction mean of 3.7.

d. The mean job satisfaction of the sample of highly trained employees at IM-ANX is significantly lower than the overall corporation satisfaction mean of 3.7.

49. The above results seem like good news for IM-ANX. The significance test indicates that the difference in job satisfaction, which was _____, is probably due to sampling error and not a systematic problem created by IM-ANX's employment environment.

a. 3.4000

b. −1.126

c. −.3000

d. .268

50. You need to compute the effect size for this significance test by hand; SPSS does not do this automatically. You can find the two values you need to compute *d* in the SPSS output. Which of the following is the correct value for *d*?
    a. −.19
    b. −.268
    c. −1.126
    d. −12.77
    e. −3.38

51. Choose the best interpretation of *d*.
    a. The sample mean is .19 above 3.
    b. The sample mean is .19 standard deviations below 3.7.
    c. The estimated statistical power is .19.

52. The significance test used a reasonable sample size ($N = 35$), and it indicated that the mean difference of _____ would occur if the null were true 13 out of 100 times. However, the mean difference is based on a sample of IM-ANX employees. If this sample did not represent the population very well, the actual difference in job satisfaction between the *population* of IM-ANX employees and the overall corporation mean satisfaction, 3.7, might actually be larger or smaller due to _____ _____.

53. Your firm's report to IM-ANX should recognize the possibility that sampling error resulted in a sample that was a poor representation of the population. This is typically done by computing a range of plausible values for the actual difference in job satisfaction, in other words, by computing a 95% confidence interval for this mean difference. Find the lower and upper boundaries for a 95% CI of this mean difference in the SPSS output.
    a. LB = −.8412, UB = .2412
    b. LB = 2.8588; UB = 3.9412
    c. LB = 1.00; UB = 5.00
    d. LB = 1.5755; UB = 3.4000

54. Choose the best interpretation of this confidence interval.
    a. There is a 95% chance that the sample mean difference is between the LB and the UB.
    b. 95% of the scores are between the LB and the UB.
    c. 95% confident that the population mean difference is between the LB and the UB.

Writing Up the Results

55. In previous activities in this book, we had you summarize results of null hypothesis significance tests without using confidence intervals. For all of the reasons mentioned in this activity, the APA recommends that researchers include confidence intervals in write-ups. Confidence intervals make us recognize how much our results might vary. Obviously, if there is a great deal of

sampling error, your results will be misleading. Confidence intervals provide a plausible range of outcomes that you should be aware of when you interpret any set of results. Now that you know about confidence intervals and why they are important, you should include them in your results summaries. The following paragraph is an excerpt of your report to IM-ANX. In the spaces, fill in the correct statistical information for a write-up of these data. When reporting numbers in APA style, round to the second decimal place.

The average job satisfaction rating of the 35 employees (M = _____, SD = _____), 95% CI [_____, _____] was not significantly lower than the overall corporation mean of 3.7, t(_____) = _____, p = _____, d = _____, 95% CI [_____, _____]. These results suggest that the employees are somewhat satisfied with their jobs, but there is room for improvement. When comparing the employees of IM-ANX to similar employees at other organizations, we find that their levels of job satisfaction were not significantly lower; however, IM-ANX should be aware that the mean was lower by .19 standard deviations. Although this is a small effect, it could have an important effect on the likelihood of employees leaving IM-ANX.

56. As a group or an individual, try to explain the meaning of each number in the above write-up. What important information does each number in the write-up convey? (You don't necessarily have to write all this out, but working through the verbal explanation by yourself or with someone may help clarify what you should have learned. Give it a try.)

Chapter 8 Practice Test

1. Why do researchers compute confidence intervals?
   a. To estimate a population parameter
   b. To determine if the outcome is likely to be due to sampling error
   c. To determine if the treatment had an effect
   d. To describe the size of the effect a treatment had on a dependent variable
2. Confidence intervals can only be computed around single means and not around mean differences.
   a. True
   b. False
3. Which of the following interpretations of a 95% confidence interval is correct?
   a. All values within a confidence interval are equally likely.
   b. If the study were repeated, the population mean would be in that confidence interval 95% of the time.
   c. A set of plausible values for a population parameter.
   d. If the mean is in the confidence interval, you can reject the null hypothesis.

SCENARIO 1: CONFIDENCE INTERVAL FOR A SINGLE MEAN

People often complain about the length and frequency of commercials on TV. A company is considering offering TV shows online without commercial interruptions for an additional cost. Before pursuing this plan, the marketing department wants to know how much people would be willing to pay for such a service. To get an estimate of the amount people will pay, they survey a random sample of 67 TV watchers and find that the mean amount is $38.65 with a standard deviation of $25.32.

4. What is the point estimate for this scenario?
5. Which table of critical values should you use to create a 95% confidence interval?
   a. One-tailed, .05
   b. One-tailed, .01
   c. Two-tailed, .05
   d. Two-tailed, .01
6. Compute the margin of error.
   a. 1.99
   b. 4.89
   c. 3.09
   d. 6.18

7. Construct a 95% confidence interval around the point estimate.
   a. LB = 36.65, UB = 40.65
   b. LB = 35.56, UB = 41.74
   c. LB = 32.47, UB = 44.63
8. Which of the following statements is the best interpretation of this confidence interval?
   a. 95% of the sample means are between the LB and the UB.
   b. 95% confident that the sample mean is between the LB and the UB.
   c. 95% confident that the population mean is between the LB and the UB.
9. Why would a larger sample size be beneficial in this case?
   a. It would make the confidence interval more precise.
   b. It would increase the standard error of the mean.
   c. It would lower the *p* value.
10. Would a 99% confidence interval be narrower or wider than the 95% confidence interval?
    a. Narrower
    b. Wider
11. The data analyst chose to compute a confidence interval to answer the research question. However, other statistical procedures could be used to answer different research questions using the same data. Match each of the following research questions to the appropriate statistical procedure.

    ____ Effect size

    ____ Confidence interval

    ____ Significance testing

    a. The company wants to know if TV watchers will pay significantly more than they are paying now if they remove commercials from TV programming.
    b. The company wants an estimate of how much TV watchers will pay for commercial-free TV.
    c. The company wants to know how effective removing commercials will be at increasing TV watchers' satisfaction with TV programming.

SCENARIO 2: CONFIDENCE INTERVAL FOR A MEAN DIFFERENCE

Scores on a standardized test of reading comprehension have a mean of 400. A principal wants to know how much his students' reading comprehension scores differ from the national average. To answer this question, he obtains reading comprehension scores from a sample of 100 students in his school and finds that the mean at his school is $M = 385$ ($SD = 98$). Compute the 95% confidence interval for the mean difference between the sample mean and a score of 400.

12. What is the point estimate for this scenario?

13. Compute the margin of error.
    a. 9.8
    b. 1.9842
    c. 18.64
    d. 19.45

14. Construct a 95% confidence interval around the point estimate.
    a. LB = 14.80; UB = 16.98
    b. LB = 5.2; UB = 24.8
    c. LB = −13.68; UB = 42.89
    d. LB = −4.45; UB = 34.45

15. Which of the following statements is the best interpretation of this confidence interval?
    a. 95% of the mean difference confidence intervals are significant.
    b. 95% confident that the sample means are within one standard error of the population mean.
    c. 95% confident that the population mean difference is between the LB and the UB.

REFERENCES

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge Academic.

CHAPTER 9

Related Samples *t* Test

LEARNING OBJECTIVES

After reading this chapter, you should be able to do the following:

- Identify when a related samples *t* test should be used
- Explain the advantages of using a related samples design over an independent samples design
- Explain the logic of the related samples *t* test
- Write null and research hypotheses using symbols and words for both one- and two-tailed tests
- Compute the degrees of freedom and define the critical region for one- and two-tailed tests
- Compute a related samples *t* test by hand (using a calculator) and using SPSS
- Determine whether you should reject the null hypothesis
- Compute an effect size (*d*) and interpret it
- Summarize the results of the analysis using American Psychological Association (APA) style
- Interpret the SPSS output for a related samples *t* test

REPEATED/RELATED SAMPLES *t* TEST

In this chapter, you will learn about a different kind of *t* test. While its computation is similar to the single-sample *t* test you learned in Chapter 7, this new *t* test goes by several different names, one for each of the experimental designs in which it can be used. For example, it can be used to compare the mean of a sample of people before and after they experience a treatment. In this situation, the *t* test is called a **repeated-measures *t* test** because *participants are measured repeatedly, once before and once after a treatment*. The repeated-measures *t* test determines if the treatment *changed* participants' scores on a dependent variable (DV). This same *t* test can also be used when researchers want to compare two groups of participants that are related to each other in some way. In this situation, the *t* test is called a **related samples *t* test** because *each person in the first sample has something in common with or is linked to someone in the second sample (i.e., the samples are related)*. For example, the first sample

might be husbands and the second sample might be their spouses. Even though the repeated-measures *t* test and the related samples *t* test have different names that arise from their different experimental designs, when it comes to computation, they are identical. In fact, the *t* test you will learn about this chapter is also sometimes called a **paired samples *t* test, matched samples *t* test, dependent samples *t* test,** or a **within-subjects *t* test.** All of these names refer to the same statistical test that you will learn in this chapter.

The related-measures *t* test is similar to the single-sample *t* test in that it compares the deviation between two means to determine if it is likely to have been created by sampling error. However, the related samples *t* test is different in that the two means it compares both come from the *same sample,* which is measured twice under different conditions. For example, if researchers wanted to test the effectiveness of a new drug intended to lower anxiety, they might take a sample of people, measure their anxiety first, then give them the new drug, and finally measure their anxiety after the drug had time to work. You might recognize this as a "pre–post" test. If the drug did not work at all, one would expect the mean anxiety level to be the same before and after taking the drug. However, if the drug worked, one would expect the mean anxiety level to be significantly lower after taking the drug than before taking it. In this situation, researchers use the same group of people to represent two different populations. The predrug sample mean represents what the population's anxiety level would be if they *did not* take the drug. The postdrug sample mean represents what the population's anxiety level would be if they *did* take the drug. Thus, the researchers are using one sample to produce two different sample means, and each sample mean estimates a different population parameter. If the deviation between the two sample means is unlikely to be created by sampling error, then the drug is assumed to have changed the anxiety level of the sample.

In some situations, researchers form pairs of people who are similar on some variable they are interested in "controlling." For example, researchers might create pairs of people who have the same anxiety scores (i.e., two people with anxiety scores of 180, two people with scores of 210, etc.). Then the researchers would randomly give one person from each "matched" pair the drug and the other a placebo. When using "matched" samples, you analyze the data as if each matched pair were really a single person. The hypotheses, critical regions, and calculations are exactly the same as those used in the previous pre–post example, but in this situation, the *t* test is frequently called a related or matched *t* test. The following hand calculation example illustrates how this type of matching procedure is done.

Reading Question

1. Researchers use the related samples *t* test to determine if _____ differ more than would be expected by sampling error.
   (a.) two related sample means
   b. a sample mean and a population mean

Reading Question

2. The related samples *t* test can be used for a study that is using
   a. a matching research design.
   b. a pre–post research design.
   (c.) both research designs.

LOGIC OF THE SINGLE-SAMPLE AND REPEATED/RELATED SAMPLES *t* TESTS

The logic of the related samples *t* test is similar to that of the single-sample *t* test. To help illustrate the similarities and differences between these two statistical tests, both formulas are given as follows:

$$\text{Single-sample } t = \frac{M - \mu}{SEM_s}.$$

$$\text{Repeated/related samples } t = \frac{M_D - \mu_D}{SEM_r}.$$

The denominators of both the single-sample *t* test and the related samples *t* test represent the typical amount of sampling error expected. The numerator of the single-sample *t* test compares a sample mean (i.e., *M*) with an expected value if the null hypothesis were true (i.e., μ). The numerator of the related samples *t* test is slightly more complicated in that rather than comparing a single-sample mean (i.e., *M*) with an expected value if the null hypothesis were true (i.e., μ), it actually compares the mean difference between two sample means (i.e., M_D; the *D* stands for difference) with the mean difference expected if the null is true (i.e., μ_D). For example, you would compare the difference between the predrug anxiety mean and the postdrug anxiety mean to the mean difference expected if the drug did not work at all (i.e., μ_D, a mean difference of 0). In fact, researchers are almost always testing to see if the observed difference is significantly different from 0. In this book, we will always use $\mu_D = 0$ because it is exceedingly rare to use a value other than 0 (SPSS does not even allow the use of other values). If you use $\mu_D = 0$, it can be eliminated from the numerator, and the *t* formula can be simplified to

$$\text{Related samples } t = \frac{M_D}{SEM_r}.$$

Another important similarity between the single-sample *t* and the related *t* tests is that if the null hypothesis for either test were true, you would expect to get an obtained *t* value close to 0. If the obtained *t* value were farther from 0 than the critical value, the null hypothesis would be rejected.

Reading Question

3. An important distinction between the single-sample *t* test and the related samples *t* test is that the _____ analyzes mean differences between two samples.

   a. single-sample *t* test

   b. related samples *t* test

Reading Question

4. The denominators of the single-sample *t* test and the related samples *t* test are both
   a. expected to be 0 if the null is true.
   b. the typical amount of sampling error expected in the study.

Reading Question

5. As with all types of *t* tests (e.g., single-sample *t* test and others), *if the null hypothesis is false,* the related samples *t* test expects an obtained *t* value that is
   a. close to 0.
   b. far from 0.

As with the single-sample *t* test, a related-measures *t* test can be either one-tailed or two-tailed. The first example is a two-tailed test.

RELATED SAMPLES *t* (TWO-TAILED) EXAMPLE

A clinical researcher wants to determine if a new drug for treating anxiety has side effects that change cholesterol levels. Cholesterol levels are normally distributed and measured on an interval/ratio scale. Six pairs of people are matched on their preexisting cholesterol levels. Then, one person in each pair is given the new anxiety drug, while the other person is given a placebo. Eighteen weeks later, the researcher measures their cholesterol levels. The clinician doesn't have any way of knowing if the anxiety drug will increase, decrease, or have no effect on cholesterol, so he used a two-tailed hypothesis test with $\alpha = .05$ to determine if the drug *changes* cholesterol levels. You've been hired to run this analyses for this study.

Step 1: Examine the Statistical Assumptions

As with the single-sample *t* test, when conducting a related samples *t* test, the *independence of data assumption* requires that participants' scores within a condition do not influence the scores of others in that same condition. Again, careful procedural controls usually produce data that are independent. When conducting a related samples *t* test, the IV must identify two groups of related scores that are measured under different conditions, and the DV must be measured on an interval or ratio scale (*appropriate measurement of variables assumption for related* t *test*). The IV in this study identifies two related groups matched on their initial cholesterol levels and then randomly assigned to two different conditions. One group took the antianxiety drug for 18 weeks while the other group took a placebo. So, the IV identifies the two conditions of drug versus placebo. The DV is participants' cholesterol level after the drug or placebo was taken for 18 weeks. Cholesterol levels are measured on an interval/ratio scale. Both the IV and the DV satisfy the appropriate measurement of variables assumption for the related samples *t* test. The *normality assumption* for a related samples *t* test requires that the distribution of *sample mean differences* be normally shaped. We will describe the distribution of sample mean differences later, but essentially the central limit theorem also applies to the distribution of *sample mean differences* so it will have a normal shape as long as the original population of mean differences has a normal shape or if the sample size is sufficiently large (i.e., greater than 30). This assumption is met because the population of cholesterol levels is known to have a normal shape, so the population of mean differences is also likely to have a normal shape. Although the normality assumption will be met in this case, in a real study, it is never a good idea to use only six pairs of participants to represent an entire population.

We use this small sample size so the example will be easier for you to follow. When you do research, it will usually be necessary to have a larger sample size. The related samples *t* test does not have a homogeneity of variance assumption. Given that all three assumptions of the related *t* test are satisfied, you can progress to Step 2.

Reading Question

6. You use a related samples *t* statistic when (choose two)
   a. the IV defines two independent samples, and the DV is measured on an interval/ratio scale.
   b. the IV defines two matched samples, and the DV is measured on an interval/ratio scale.
   c. the IV defines one sample measured under two different conditions, and the DV is measured on an interval/ratio scale.
   d. the IV defines one sample and the DV is measured on an interval/ratio scale, and you do not know the population standard deviation.
   e. the IV defines one sample and the DV is measured on an interval/ratio scale, and you do know the population standard deviation.

Reading Question

7. Which of the following common statistical assumptions is not an assumption of the related samples *t* test?
   a. Data independence
   b. Appropriate measurement of the IV and the DV
   c. The distribution of sample mean differences must have a normal shape
   d. Homogeneity of variance

Step 2: State the Null and Research Hypotheses Symbolically and Verbally

In this case, you are not sure what effect the antianxiety drug has on cholesterol levels, so you correctly choose to do a two-tailed hypothesis test. Your research hypothesis states that the drug does affect cholesterol levels. This hypothesis is nondirectional in that it does not specify if the drug has a positive or negative effect, just that it will have some effect. As will always be the case, your null hypothesis states the opposite of the research hypothesis—namely, that the drug does not affect cholesterol levels in any way.

In this example, you are using pairs of people who are matched on their preexisting cholesterol levels. You treat each matched pair as if they were a single participant. One person in each matched pair will get the drug, while the other will get the placebo. Those who get the drug help create a mean that represents the population's cholesterol if everyone took the drug. Those who get the placebo help create a mean representing the population's cholesterol if everyone took the placebo. The related samples *t* test does not directly compare these means. Instead, the related samples *t* test uses the difference scores from each matched pair of participants. You must compute these difference scores by subtracting each pair's placebo cholesterol score from their drug cholesterol score. You will see this in Step 4. The related samples *t* test determines if the mean of these difference scores is significantly different from 0. If the drug has no effect on cholesterol, the mean difference for all the matched pairs should be close to 0. The symbolic notation for the mean of the difference scores is μ_D. Thus, the null hypothesis

is that the mean of the difference scores is equal to 0 ($\mu_D = 0$), and the research hypothesis is that the mean of the difference scores is not equal to 0 ($\mu_D \neq 0$). The null and research hypotheses are shown in Table 9.1.

Reading Question

8. In most research situations in which a related samples *t* test is used, the μ_D expected if the null is true is
   a. 0.
   b. 1.
   c. M_D.

Table 9.1 Symbolic and Verbal Representations for Two-Tailed Research and Null Hypotheses for a Related Samples *t* Test

	Symbolic	*Verbal*	*Mean Difference Created by*
Research hypothesis (H_1)	H_1: $\mu_D \neq 0$	The mean cholesterol level in the population of people who take the placebo is different from the mean cholesterol level in the population of people who take the antianxiety drug.	The drug having a different effect on cholesterol than the placebo
Null hypothesis (H_0)	H_0: $\mu_D = 0$	The mean cholesterol level in the population of people who take the placebo is not different from the mean cholesterol level in the population of people who take the antianxiety drug.	Sampling error

Reading Question

9. Which of the following best represents the null hypothesis for a two-tailed related samples *t* test?
   a. $\mu_D \neq 0$.
   b. $\mu_D = 0$.
   c. $\mu_1 = \mu_2$.

Reading Question

10. μ_D is the symbolic notation for
    a. the null hypothesis.
    b. the mean of the difference scores from the population.

Step 3: Compute the Degrees of Freedom and Define the Critical Region

This research scenario uses a matched design. This means that you are using a pair of people who are matched on their cholesterol scores as if they were just one person. Therefore, when computing the

degrees of freedom (*df*) for a matched design, *N* is the number of *paired scores*. In this case, *N* would be 6, and the *df* would be

$$df = (N-1) = (6-1) = 5 .$$

To determine the critical value, you use the same table of critical *t* values we have used for all other *t* tests and find it to be 2.5706 when $\alpha = .05$. This means that the two-tailed critical regions are $t < -2.5706$ and $t > +2.5706$.

Step 4: Compute the Test Statistic (Related Samples *t* Test)

4a. Compute *D* for Each Participant/Matched Pair

The first step in computing the related samples *t* statistic is computing the difference score (i.e., *D*) for each pair of scores. You must compute the difference in the same way for each pair of scores. In this case, the difference score was computed as Drug Score minus Placebo Score. The *D* for each pair of scores is computed in Table 9.2. *D* is the difference between the two scores for each participant pair.

Table 9.2 Computation of *D* in a Related Samples *t* Test

Pair	*Placebo*	*Drug*	D *(Drug – Placebo)*
A	180	188	8
B	200	201	1
C	190	197	7
D	170	174	4
E	210	215	5
F	195	194	−1

Reading Question

11. Difference scores will always be positive.
    a. True
    b. False

Reading Question

12. The first step in computing a related samples *t* is to compute difference scores (*D*s) by
    a. subtracting the mean for Group 1 from the mean for Group 2.
    b. computing the *D* for each set of paired scores.

4b. Compute the Observed Mean Difference (M_D)

Next compute the numerator of the t (i.e., the mean of the difference scores) by adding the Ds and dividing by the number of Ds (N):

$$M_D = \frac{\sum D}{N} = \frac{24}{6} = 4.$$

4c. Compute the Average Mean Difference Expected Due to Sampling Error

Now compute the typical mean difference expected due to sampling error. At first glance, the sum of squares (SS) formula shown below appears different from the SS formula we used in previous chapters. The difference is only minor, however. The previous SS formula used the $\sum X$ and $\sum X^2$. The following SS formula uses $\sum D$ and $\sum D^2$ because the related samples t test analyzes the Ds, not the scores themselves. The computational process is identical; you just use the difference scores (i.e., Ds) instead of scores (Xs):

$$SS_D = \sum D^2 - \frac{(\sum D)^2}{N} = 156 - \frac{(24)^2}{6} = 60.$$

$$SD_D = \sqrt{\frac{SS_D}{N-1}} = \sqrt{\frac{60}{5}} = \sqrt{12} = 3.464.$$

$$SEM_r = \frac{SD_D}{\sqrt{N}} = \frac{3.464}{\sqrt{6}} = 1.414.$$

This value of 1.414 is the typical sampling error for this study. It is the expected difference between these sample means due to sampling error.

Reading Question

13. The related samples t test analyzes _____ rather than _____.
    a. deviations scores; difference scores
    b. raw scores; difference scores
    c. difference scores; raw scores

Reading Question

14. Which of the following values is a measure of sampling error?
    a. SS_D
    b. SD_D
    c. SEM_r

4d. Compute the Test Statistic (Related Samples t Test)

The obtained t value for the related samples t test is computed as follows:

$$t = \frac{M_D}{SEM_r} = \frac{4}{1.414} = 2.83.$$

The obtained *t* value of 2.83 is farther from 0 than the critical value of +2.5706; therefore, it is in the positive critical region. Thus, you should reject the null hypothesis and conclude that people who took the antianxiety drug had higher cholesterol levels than those who took the placebo. This value of 2.83 indicates that the difference between the sample means was 2.83 times greater than would be expected by chance if the null hypothesis is true. This obtained difference is unlikely to be due to sampling error.

Reading Question

15. If the obtained *t* value (i.e., in this case 2.83) is farther from 0 than the critical value, the difference between the two means is
    a. likely to be due to sampling error.
    b. not likely to be due to sampling error.

Step 5: Compute an Effect Size and Describe It

Computing the effect size (*d*) is done the same way as for the single-sample *t* test.[1] Specifically,

$$d = \frac{\text{Observed deviation between the means}}{\text{Standard deviation of difference scores}} = \frac{M_D}{SD_D} = \frac{4}{3.464} = 1.15 .$$

When computing a *d* for a related samples *t* test, the denominator is the standard deviation of the *D* scores (SD_D).

The same effect size cutoffs are used as with the single-sample *t* test. Specifically, the effect size is small if *d* is close to .2, medium if it is close to .5, and large if it is close to .8. You should always use the absolute value when determining the size of the effect. In this case, the *d* of 1.15 is a large effect size. The difference in cholesterol scores between the placebo and drug conditions is 1.15 times larger than the standard deviation of *D* scores. This suggests that the new antianxiety drug has a large effect on cholesterol levels.

Reading Question

16. To compute the effect size, you divide the observed deviation between the means by
    a. the standard error.
    b. the standard deviation of the difference scores.
    c. the mean standard deviation of the scores.

Step 6: Interpreting the Results of the Hypothesis Test

The following sentences summarize the results of this related samples *t* test. You will need to compute the means for the drug and placebo groups by hand or using SPSS to include them in your report of the results.

1 Many researchers use this formula for computing *d*, but there are other options. For example, Cumming (2012) recommends dividing the mean of the differences scores by the average of the standard deviations rather than the standard deviation of the difference scores. Either calculation is acceptable as long as you tell the reader how you computed the effect size. In this book, we are always using the same calculation, and so we do not say repeatedly how it was done, but when you present data, you should state how *d* was calculated.

People who took the antianxiety drug had substantially higher cholesterol levels (M = 194.83, SD = 13.64) than people who took the placebo (M = 190.83, SD = 14.29), $t(5) = 2.83$, $p < .05$, $d = 1.15$. The drug raised cholesterol levels by more than 1 standard deviation, suggesting the drug has a detrimental effect on cholesterol levels. However, the sample included just six pairs of people, and so the study should be replicated with a larger sample size.

Reading Question

17. The *effect size* in this study indicates that the antianxiety drug raised cholesterol levels of the participants
    a. by 1.15 standard deviations, which is a very large effect.
    b. 1.15 times more than would be expected by chance.

In the previous example, you did not know if the anxiety drug would have an effect on cholesterol levels, and so she chose to do a two-tailed test. In the next example, you are testing the effect of the antianxiety drug on anxiety. Because you clearly expect the drug to reduce anxiety, you should do a one-tailed test.

RELATED SAMPLES *t* TEST (ONE-TAILED) EXAMPLE

A new drug for treating anxiety has been developed. Obviously, it is expected to lower the anxiety scores of people who take it. A psychiatrist recruits six people with high anxiety scores for an evaluation study of the new drug. All six volunteers complete an anxiety inventory before taking the drug and then again after taking the drug for 1 month. Their pre- and postanxiety scores are displayed in Table 9.4. Scores on the anxiety inventory are normally distributed and measured on an interval/ratio scale. You are hired to help run this study. You correctly decide to use a one-tailed related samples t test with an alpha of .05 to determine if this new drug lowers anxiety scores.

Step 1: Examine the Statistical Assumptions

As was the case for the previous study, the participants' responses within each condition are independent of each other, the IV identifies two conditions (i.e., before vs. after taking the drug) and the DV is measured on an interval/ratio scale, and the distribution of sample mean differences is likely to have a normal shape. Again, it is generally better to have a sample size larger than is used here, but this small sample will work well as an example. Given that all three assumptions are met, it is appropriate to move on to the Step 2.

Step 2: State the Null and Research Hypotheses Symbolically and Verbally

The null and research hypotheses are given in Table 9.3.

Your hypotheses depend on how you compute the difference scores. In this case, the difference scores were computed as After Drug minus Before Drug. The research hypothesis predicts that anxiety scores will be higher before taking the drug and lower after taking the drug. Therefore, *if the research hypothesis is correct* and the difference scores are computed as After Drug (lower anxiety) – Before Drug (higher anxiety), the mean difference score (μ_D) would be negative.

Table 9.3 Symbolic and Verbal Representations for One-Tailed Research and Null Hypotheses for a Related *t* Test

	Symbolic	*Verbal*	*Mean Difference Created by*
Research hypothesis (H_1)	H_1: $\mu_D < 0$	The population's mean anxiety score is lower after taking the antianxiety drug than before taking it.	The antianxiety drug lowering cholesterol
Null hypothesis (H_0)	H_0: $\mu_D \geq 0$	The population's mean anxiety score is not lower after taking the antianxiety drug than before taking it.	Sampling error

Reading Question

18. Which of the following *could* represent the null hypothesis for a one-tailed related samples *t* test?
    a. $\mu_D \leq 0$
    b. $\mu_D \leq 0$
    c. $\mu_1 = \mu_2$
    d. $\mu_D \geq 0$

Step 3: Compute the Degrees of Freedom and Define the Critical Region

The *df* is computed identically for one- and two-tailed tests:

$$df = N - 1 = 6 - 1 = 5 .$$

In this case, the critical region is $t < -2.0150$.

Reading Question

19. When computing the *df* for a related samples *t* test, the *N* in the formula is the
    a. number of scores.
    b. number of pairs of scores.

Step 4: Compute the Test Statistic (Related Samples *t* Test)

4a. Compute *D* for Each Participant/Matched Pair

As mentioned earlier, the related samples *t* test analyzes difference scores (*D*s) rather than raw scores. The difference score for each participant is computed in Table 9.4. All difference scores were computed by subtracting After Drug – Before Drug scores.

Table 9.4 Computing *D* for a Related *t* Test

Volunteer	*Before*	*After*	D *(After – Before)*
A	22	19	−3
B	22	19	−3
C	23	18	−5
D	26	25	−1
E	23	20	−3
F	25	21	−4

Reading Question

20. When computing a related samples *t* test, you must remember that all of the analyses are done on
    a. the difference between the paired scores for each participant (i.e., *D*).
    b. the raw scores of each participant.

4b. Compute the Observed Mean Difference (M_D)

The mean difference score (i.e., the numerator of the related samples *t* test) is computed as follows:

$$M_D = \frac{\sum D}{N} = \frac{-19}{6} = -3.167.$$

4c. Compute the Average Mean Difference Expected Due to Sampling Error

The three computational steps for computing the expected sampling error are given as follows:

$$SS_D = \sum D^2 - \frac{(\sum D)^2}{N} = 69 - \frac{(-19)^2}{6} = 8.833.$$

$$SD_D = \sqrt{\frac{SS_D}{N-1}} = \sqrt{\frac{8.833}{5}} = 1.329.$$

$$SEM_r = \frac{SD_D}{\sqrt{N}} = \frac{1.329}{\sqrt{6}} = .543.$$

This value of .543 is the typical expected sampling error. It is the size of the anxiety difference between the After Drug and Before Drug means that is expected due to sampling error.

4d. Compute the Test Statistic (Related Samples *t* Test)

The obtained *t* value for the related samples *t* is computed as follows:

$$t = \frac{M_D}{SEM_r} = \frac{-3.167}{.543} = -5.83.$$

The obtained *t* value of −5.83 is farther from 0 than the critical value of −2.0150; therefore, it is in the negative critical region. Thus, you should reject the null hypothesis and conclude that the anxiety scores were lower after taking the drug than before taking the drug.

Reading Question

21. The obtained *t* value in this study indicates that
    a. the deviation between the sample means was 5.83 times larger than was expected due to sampling error.
    b. the null hypothesis should be rejected.
    c. both a and b are correct.

Step 5: Compute an Effect Size and Describe It

The effect size for this study is computed as follows:

$$d = \frac{\text{Observed deviation between the means}}{\text{Standard deviation}} = \frac{M_D}{SD_D} = \frac{-3.167}{1.329} = -2.38.$$

A *d* of −2.38 is a large effect size. This means that the difference in anxiety levels between the Before Drug and After Drug treatment conditions is 2.38 times larger than the standard deviation of difference scores. This suggests that the new antianxiety drug is very effective at lowering anxiety levels.

Reading Question

22. The *effect size* in this study indicates that the drug lowered the anxiety of the participants
    a. 2.38 times more than would be expected by chance.
    b. by 2.38 standard deviations, which is a very large effect.

Step 6: Interpreting the Results of the Hypothesis Test

The following sentences summarize the results of this study:

People had lower anxiety levels ($M = 20.33$, $SD = 2.50$) after taking the drug than they had prior to taking the drug ($M = 23.50$, $SD = 1.64$), $t(5) = -5.83$, $p < .05$ (one-tailed), $d = -2.38$. The reduction in anxiety scores was quite large, more than 2 standard deviations. The sample size in this study was very small. A larger study should be done before any conclusions are drawn about the drug.

Reading Question

23. When writing the reporting statement of the results, the *p* value is written as less than .05 (i.e., $p < .05$) because
    a. the obtained *t* value was in the critical region.
    b. the null hypothesis was not rejected.

STATISTICAL RESULTS, EXPERIMENTAL DESIGNS, AND SCIENTIFIC CONCLUSIONS

The result of the above significance test seems to imply that the antianxiety drug works, and the effect size seems to imply that it is very effective. However, all statistical results must be interpreted cautiously, particularly paying close attention to the study's experimental design. In this study, the significance test indicates that something other than sampling error probably created the observed reduction in anxiety. Therefore, it may be tempting to conclude that the drug caused the reduction, but there is at least one other potential cause. In this pre–post design, participants' anxiety scores were measured before and after taking the drug; these measurements were separated by 1 month. One possible explanation is that people naturally get better with the passage of time. Participants might have gotten better without any treatment. Therefore, the passage of time is a *confound* in this study. It is also correct to say that the passage of time and the drug treatment are *confounded* in this study because it is impossible to know whether the passage of time, the drug, or some combination of both caused the reduction in anxiety. The lesson here is that you must consider potential confounding variables carefully before you interpret *any* statistical result. Ideally, researchers should use experimental designs that control for potential confounding variables. For example, researchers could control for improvement with time by using a control group of different participants who only received a placebo drug. If the placebo treatment did not reduce anxiety while the drug treatment did, the researchers could be pretty confident that the drug caused the anxiety reduction and not the passage of time. The key point is that a statistical result (i.e., rejecting a null or a large effect size) is only one portion of a convincing scientific argument. If a study has a confound, the statistical results are not very informative. Statistical results and experimental design are equally important to the scientific process. You will learn how to use experimental designs to control potential confounding variables in a research methods course.

Reading Question

24. When interpreting the statistical results of any study,
    a. you should consider if a confounding variable might have affected the results.
    b. you should recognize that the experimental design is just as important to the scientific conclusion as the statistical results.
    c. both of the above are correct.

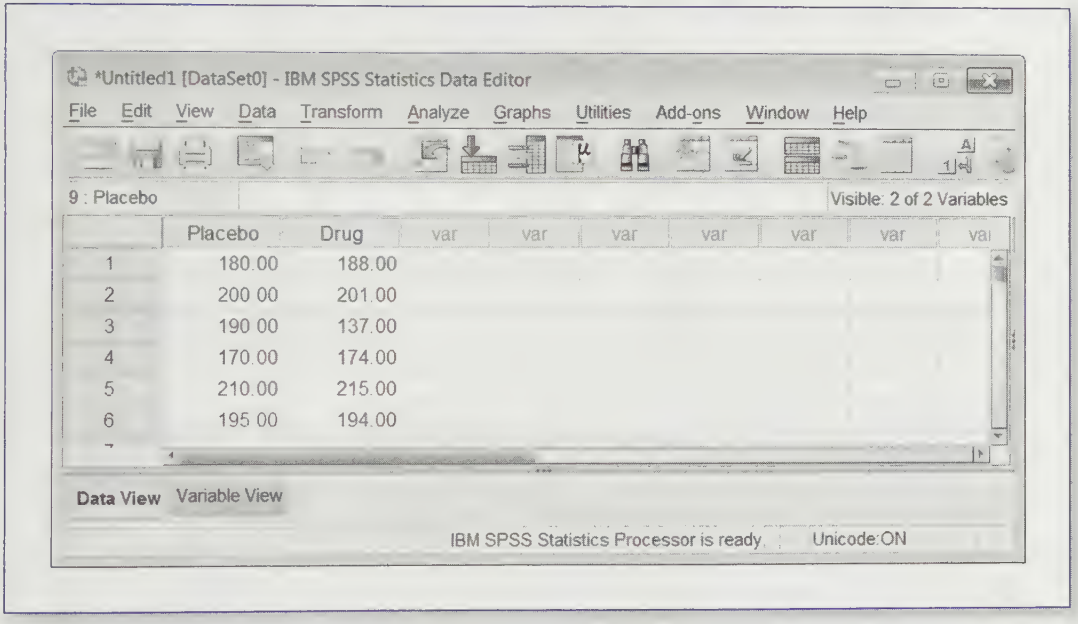
SPSS

Data File

We used the data from the first example (i.e., the effect of an antianxiety drug vs. placebo on cholesterol levels) to demonstrate how to run a related samples t test. Related measures data are entered into two columns in the data file. One column should have the placebo cholesterol levels, and the other column

should have the drug cholesterol levels. When entering matched data, you must enter the data for matched participants on the same row in the data file. Likewise, when entering related samples data, in which the same people have been measured twice, you must enter the data for each participant on the same row in the data file. After you have entered them, the data from the first example problem should look like what is shown in Figure 9.1.

Figure 9.1 SPSS Screenshot of the Data Entry Screen



Computing a Related Samples *t* Test

- Click on the Analyze menu. Choose Compare Means and then Paired Samples *t* Test (see Figure 9.2).
- Move both independent variable conditions (in this case, Drug and Placebo) into the Paired Variables box. In some versions of SPSS, you have to move them into the box at the same time (see Figure 9.3).
- Click on the OK button.

Figure 9.2 SPSS Screenshot of Choosing a Paired Samples *t*

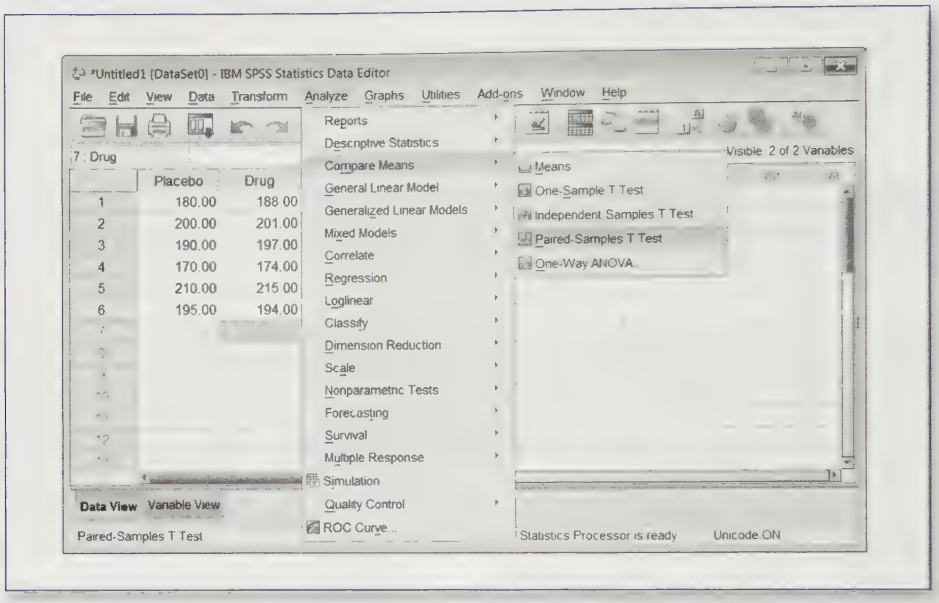
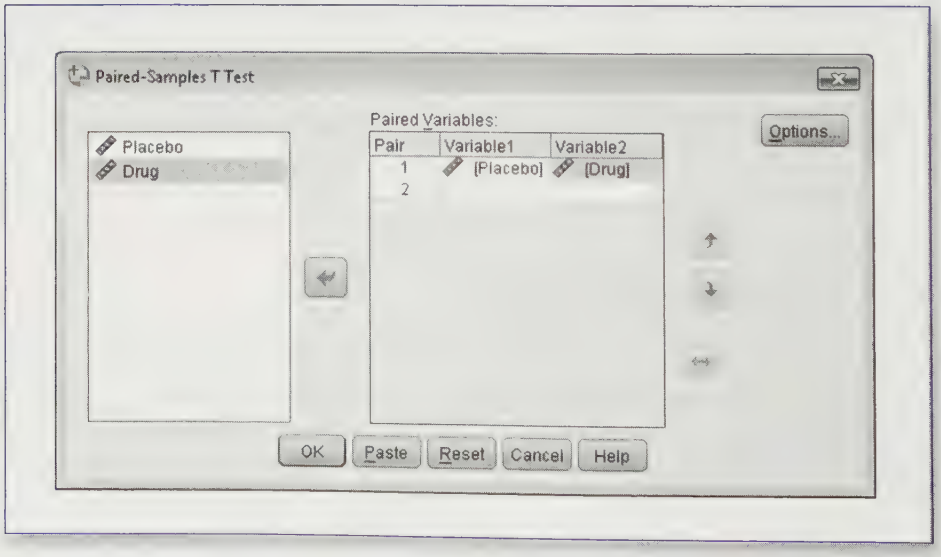


Figure 9.3 SPSS Screenshot of Selecting the Paired Variables



Output

Figure 9.4 Annotated SPSS Output for a Related *t* Test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Drug	194.8333	6	13.64429	5.57026
	Placebo	190.8333	6	14.28869	5.83333

Mean: Sample means (*M*)

N: Sample sizes

Std. Deviation: Standard deviations for each treatment (*SD*)

Std. Error Mean: Standard errors of the means; Computed as $SD_D/\sqrt{N}$

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Drug & Placebo	6	.970	.001

Paired Samples Correlations: This table is not needed to interpret a *t* test.

Pair 1: Drug–Placebo: Tells you which mean was subtracted from the other. In this case, the Placebo mean was subtracted from the Drug mean.

95% Confidence Interval: We are 95% confident that the actual difference between the sample and population means is between the lower and upper values.

Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	*t*	df	Sig. (2-tailed)
Pair 1	Drug - Placebo	4.00000	3.46410	1.41421	.36465	7.63535	2.828	5	.037

Mean: Mean of the difference scores (M_D); difference between the two sample means

Std. Deviation: Standard deviation of the difference scores (SD_D)

Std. Error Mean: Standard error of the difference scores (SEM_Y); the denominator of the *t*

t: Obtained *t* value

$$t = \frac{M_D}{SEM_Y}$$

df: Degrees of freedom ($N - 1$)

Sig. (2-tailed): Two-tailed *p* value; the probability of obtaining the obtained *t* value or a more extreme *t* value if the null hypothesis is true;

For a one-tailed test, divide this value by 2; reject H_0 if $p < \alpha$.

Reading Question

25. Use the "Paired Samples Statistics" output to determine the mean and standard deviation for the drug condition. The mean and standard deviation for the drug condition were
   a. 194.83 and 13.64, respectively.
   b. 190.83 and 14.29, respectively.

Reading Question

26. When you enter data for a related samples *t* test, you should
   a. have the paired scores on the same row of the spreadsheet.
   b. have as many columns as you have scores.

Reading Question

27. Use the "Paired Samples Test" output to determine which of the following was the numerator of the *t* test.
   a. 4.0000
   b. 3.46410
   c. 1.41421
   d. 2.828
   e. .037

Reading Question

28. Use the "Paired Samples Test" output to determine which of the following was the denominator of the *t* test.
   a. 4.0000
   b. 3.46410
   c. 1.41421
   d. 2.828
   e. .037

Reading Question

29. Use the "Paired Samples Test" output to determine which of the following was the obtained value for the *t* test.
   a. 4.0000
   b. 3.46410
   c. 1.41421
   d. 2.828
   e. .037

Reading Question

30. Use the "Paired Samples Test" output to determine which of the following values is the *p* value for this *t* test.
   a. 4.0000
   b. 3.46410
   c. 1.41421
   d. 2.828
   e. .037

OVERVIEW OF THE ACTIVITY

In Activity 9.1, you will work through all of the steps of hypothesis testing for a one-tailed hypothesis test and a two-tailed hypothesis test. You will do the calculations using a calculator and SPSS. The activity also asks you to compare and contrast the results of significance tests and effect sizes, as well as compare the results of different studies. Finally, you will also read research scenarios and determine the appropriate statistic for each situation. In Activity 9.2, you will review hypothesis testing with the related samples *t* and work with confidence intervals.

Activity 9.1: Hypothesis Testing With the Related Samples *t* Test (or Dependent *t* Test)

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Explain how the direction in which one computes the difference scores affects the sign of the obtained *t* value of a related samples *t* test
- Determine whether a one- or a two-tailed significance test is appropriate
- State the null and research hypotheses for a dependent *t* test in words and symbols
- Compute a related *t* test by hand and using SPSS
- Interpret and summarize the results of a related samples *t* test
- Determine which statistic should be used to answer different research questions

RESEARCH SCENARIO 1 (ONE-TAILED)

In this scenario, you want to determine if viewing a documentary on the dangers of using a cell phone while driving would *decrease* the frequency of cell phone use while driving. Prior to watching the video, 15 participants answered several questions measuring the extent to which they agreed with statements like "I am likely to use my cell phone while driving," using a Likert response scale, where 1 = *strongly disagree* and 7 = *strongly agree*. Participants' answers to all of the questions were totaled and averaged to create an overall score of likely cell phone use while driving. Previous research found that these averaged scores are normally distributed and measured on an interval/ratio scale. The teenage drivers then answered the same questions again after watching the video. The data are listed on the following page.

1. Match the assumption to the fact that suggests that assumption was met.

   _____ Independence

   _____ Appropriate measurement of the IV and the DV

   _____ Normality

Before	*After*
5	5
6	5
4	4
4	4
6	5
6	5
5	6
7	6
7	7
7	6
5	4
6	6
3	3
2	2
5	5

a. The population of DV scores has a normal shape.

b. Each participant answers the questions with no one else around.

c. The prevideo and postvideo conditions are clearly defined. Responses to the questionnaire are measured on an interval/ratio scale.

2. In this study, the same participants are measured twice, and so the same people are in both conditions. Why doesn't this violate the assumption of independence?

   a. The assumption is that there is independence within conditions, not between conditions.

   b. The assumption of independence is not important for a related measures design.

The Direction of Difference Scores

3. If you assume that the video has the desired effect on cell phone use while driving and you compute the difference scores as $M_{before} - M_{after}$, would you expect negative or positive difference scores?

   a. Positive

   b. Negative

4. If you assume that the video has the desired effect on cell phone use while driving and you compute the difference scores as $M_{after} - M_{before}$, would you expect negative or positive difference scores?

   a. Positive

   b. Negative

Deciding on the Alpha Value and the Number of Critical Regions

5. Should you use a one-tailed test or a two-tailed test?

   a. One-tailed. You have a clear prediction about which mean will be higher and are only interested in a program that might decrease the likelihood of using a cell phone when driving.

   b. Two-tailed. The documentary will be equally effective for your intended purpose if it increases or decreases cell phone use while driving.

6. You need to decide if you should use an alpha of .01 or .05. Which alpha level would have a higher Type I error rate?

   a. .05

   b. .01

7. Which alpha level would have a higher Type II error rate?

   a. .05

   b. .01

8. Which alpha level would have more statistical power?
   a. .05
   b. .01

9. In this study, the researchers want to maximize their statistical power. They are not overly concerned with making a Type I error because there are few costs associated with a Type I error. Which alpha level should the researcher choose?
   a. .05
   b. .01

Null and Research Hypotheses

10. Write H_0 next to the symbolic notations for the null hypothesis and H_1 next to the research hypothesis. Regardless of what you wrote earlier, do a one-tailed test. A one-tailed test is probably the best choice because you expect the intervention to reduce cell phone use while driving. Remember, these hypotheses make predictions about the "mean difference" expected between the responses provided before the video and after the video. Make your choices based on knowing that we will compute the difference scores as $M_{before} - M_{after}$.
    a. _____ $\mu_D = 0$
    b. _____ $\mu_D \neq 0$
    c. _____ $\mu_D > 0$
    d. _____ $\mu_D < 0$
    e. _____ $\mu_D \geq 0$
    f. _____ $\mu_D \leq 0$

11. Write H_0 next to the verbal description of the null hypothesis and H_1 next to the research hypothesis.
    a. _____The population of students who watch the video will be less likely to say that they will use their cell phone while driving after watching the video than before watching the video.
    b. _____The population of students who watch the video will be more likely to say that they will use their cell phone while driving after watching the video than before watching the video.
    c. _____The population of students who watch the video will not be less likely to say that they will use their cell phone while driving after watching the video than before watching the video.
    d. _____The population of students who watch the video will not be more likely to say that they will use their cell phone while driving after watching the video than before watching the video.

Defining the Critical Region

12. Compute the *df* and locate the critical region. Use an alpha of .05. Draw an estimated *t* distribution and locate the critical region on that curve.

13. How will you decide if you should reject the null based on the critical region you drew earlier?
    a. Reject H_0 if the obtained (i.e., computed) t value is outside of the critical region.
    b. Reject H_0 if the obtained (i.e., computed) t value is in the critical region.

Computing the Dependent *t* Test

14. Compute the mean of the difference scores (i.e., M_D or $M_{before} - M_{after}$).

15. Compute the standard error of the difference scores $\left(\text{i.e.,} SEM_r = \frac{SD_D}{\sqrt{N}}\right)$.

16. You should have found the standard error of the mean difference is .159. Which of the following is the best interpretation of the value?
    a. The sample mean is .159 away from the population mean with this sample size.
    b. The pretest mean is .159 away from the posttest mean due to sampling error.
    c. The typical mean difference between sample means of this size will be .159 simply due to sampling error.
    d. All possible sample mean differences are within .159 of the actual population mean differences.

17. Compute the obtained t value for this study $\left(\text{i.e.,} t = \frac{M_D}{SEM_r}\right)$.

Evaluating the Likelihood of the Null Hypothesis

18. Determine whether or not you should reject the null hypothesis.
    a. Reject H_0
    b. Fail to reject H_0

Computing the Effect Size

19. Compute the effect size of this study $\left(\text{i.e.,} d = \frac{M_D}{SD_D}\right)$.

20. Interpret the effect size as small, small to medium, and so on.

Summarizing the Results

21. Summarize the results of this study using APA style by filling in the blanks below. You have not yet computed the means and standard deviations for before and after the video. You will need to do this to complete the summary statement.

    People indicated that they were significantly less likely to use a cell phone while driving after watching the video (M = _____, SD = _____) than before watching the video (M = _____, SD =_____), t(_____) = _____, $p < .05$ (one-tailed), d = _____.

    (Note that we wrote $p < .05$ because we rejected the null and do not have the exact p value.)

RESEARCH SCENARIO 2 (TWO-TAILED)

After reviewing the results of the previous study, you wonder what effect watching a video on the dangers of cell phone use while driving might have on the use of hands-free cell phones while driving. You conduct a study similar to the one described above. Ten drivers rated their agreement with statements like "I am likely to use a hands-free cell phone while driving," using a Likert scale, where 1 = *strongly disagree* and 7 = *strongly agree*. Each participant's average responses to these questions both before and after watching a video are at right:

Before	*After*
6	5
5	5
4	6
5	6
6	5
7	5
5	7
7	6
7	7
5	7

Null and Research Hypotheses

22. You are not sure what effect the video might have on the use of hands-free cell phones. It is possible that the video will lead to an overall decrease in all cell phone use, but it is also possible that people will respond to the video by choosing to use their cell phones with hands-free devices. Thus, a two-tailed test is more appropriate than a one-tailed test. Which of the following best represents the *null* hypothesis?

    a. $\mu_D = 0$
    b. $\mu_1 = \mu_2$
    c. $\mu_D \neq 0$
    d. $\mu_1 \neq \mu_2$

23. Which of the following best represents the research hypothesis?

    a. $\mu_D = 0$
    b. $\mu_1 = \mu_2$
    c. $\mu_D \neq 0$
    d. $\mu_1 \neq \mu_2$

Defining the Critical Region

24. Compute the *df* and locate the critical region. Use an alpha of .05, two-tailed. If it helps, draw the *t* distribution and locate the critical region on that curve.

25. The *t* values in the critical regions are best described as
    a. values that are very unlikely if the null hypothesis is true.
    b. values that are very likely if the null hypothesis is true.
    c. values that are very unlikely if the research hypothesis is true.

Computing the Dependent *t* Test

26. Compute the mean of the difference scores ($M_{before} - M_{after}$).

27. What do you expect the mean of the difference scores to equal if the null hypothesis is true?

28. Compute the standard error of the difference scores.

29. Compute the obtained *t* value for this study.

Evaluating the Likelihood of the Null Hypothesis

30. Determine whether you should reject or fail to reject the null hypothesis.
    a. Reject H_0
    b. Fail to reject H_0

Computing the Effect Size

31. Compute the effect size estimate of this study.

32. Describe the size of the effect as small, small to medium, and so on.

Summarizing the Results

33. Read and evaluate the following APA-style summary of the hands-free cell phone study. Then determine if there is an error or omission and, if so, identify the problem. (Select all that apply.)

    The hands-free cell phone use ratings after watching the video were not significantly different from the ratings before watching the video.

    a. There are no errors or omissions in the above APA summary.
    b. The means and the standard deviations for each condition are missing.
    c. The statistical *t* test information is missing.
    d. The sentence implies that the means are significantly different when they are not significantly different.

34. Read and evaluate the following APA-style summary of the hands-free cell phone study. Then determine if there is an error or omission and, if so, identify the problem. (Select all that apply.)

    The hands-free cell phone use ratings after watching the video ($M = 5.9$, $SD = .88$) were higher than they were before watching the video ($M = 5.7$, $SD = 1.06$), $t(9) = -.43$, $p > .05$ (two-tailed), $d = .14$.

    a. There are no errors or omissions in the above APA summary.
    b. The means and standard deviations for each condition are missing.
    c. The statistical *t* test information is missing.
    d. The sentence implies that the means are significantly different when they are not significantly different.

35. Read and evaluate the following APA-style summary of the hands-free cell phone study. Then determine if there is an error or omission and, if so, identify the problem. (Select all that apply.)

    The hands-free cell phone use ratings after watching the video ($M = 5.9$, $SD = .88$) were not significantly different than they were before watching the video ($M = 5.7$, $SD = 1.06$), $t(9) = -.43$, $p > .05$ (two-tailed), $d = .14$.

    a. There are no errors or omissions in the above APA summary.
    b. The means and standard deviations for each condition are missing.
    c. The statistical *t* test information is missing.
    d. The sentence implies that the means are significantly different when they are not significantly different.

SPSS FOR RELATED SAMPLES *t* TESTS

Enter the data for Research Scenario 1 into SPSS. You should have two columns of data: one for the scores before the video and another column for the scores after watching the video. After the data are entered, compute a related samples *t* by clicking on the Analyze menu. Choose "Compare Means" and then "Paired Samples *t* test." Move the before and after scores into the Variable 1 and Variable 2 boxes, respectively.

36. Record the means and standard deviations for the responses before and after watching the videos.

M_{before} = _____, SD_{before} = _____.

M_{after} = _____, SD_{after} = _____.

37. In the SPSS output, find and record the value of the numerator of the *t* statistic. Verify that it is the same as what you computed by hand.

38. In the SPSS output, find and record the value of the denominator of the *t* statistic. Verify that it is the same as what you computed by hand.

39. In the SPSS output, find and record the value of the *t* statistic. Verify that it is the same as what you computed by hand.

40. In the SPSS output, find the *p* value for this statistic. What is the *p* value?

41. What does the *p* value mean?
    a. The probability of making a Type II error
    b. The probability of obtaining a mean difference of .33 or greater if the null hypothesis is true is .0275 (one-tailed)

42. How do you decide if you should reject or fail to reject H_0 based on the *p* value?
    a. Reject the null hypothesis if the *p* is less than the alpha.
    b. Reject the null hypothesis if *p*/2 is less than the alpha.
    c. Reject the null hypothesis if the *p* value is less than the critical value.

43. Did you reject the null hypothesis?

44. Enter the data for the second research scenario into SPSS, and run a paired samples *t* test. Verify that the *t* value is the same as what you computed by hand.

45. Fill in the following blanks. The *p* value for this significance test was equal to _____. The *p* value is the probability of obtaining a *t* value of _____, or more extreme, if the _____ hypothesis is true.

46. For Research Scenario 2, you used a two-tailed test. Explain how you will decide whether to reject the null hypothesis when given a two-tailed *p* value from SPSS.
    a. Reject the null hypothesis if the *p* is less than alpha.
    b. Reject the null hypothesis if *p*/2 is less than alpha.
    c. Reject the null hypothesis if the *p* value is less than the critical value.

Putting It All Together

47. A social psychologist investigating the effects of mood on risky decision making needs to manipulate mood and is looking for a way to put participants in a negative mood. Rather than develop a new method, she turns to the literature to find ways that other researchers have induced a negative mood.

    In one study, researchers had participants fill out the Positive and Negative Affect Scale (PANAS), a questionnaire designed to assess mood, where higher numbers indicate more negative mood. The same participants then watched a 10-minute clip of a sad movie in which a young boy watches his father die as the result of a boxing match. After the movie, they filled out the PANAS again. The mean PANAS score prior to watching the movie was $M = 1.85$, while the mean score after watching the movie was $M = 3.16$, $t(54) = 6.22$, $p < .01$ (two-tailed, $\alpha = .05$), $d = .83$.

    In another study, researchers had participants fill out the PANAS before and after a mood induction procedure. In this study, negative mood was induced by giving participants negative feedback. Specifically, participants took a challenging 30-item math test and were then told that they did very poorly on the test compared with their peers. The mean PANAS score prior to the false feedback was $M = 1.99$, while the mean score after getting the negative feedback was $M = 3.89$, $t(18) = 4.32$, $p < .01$ (two-tailed, $\alpha = .05$), $d = .99$.

    Which mood induction procedure do you think the researcher should use? Explain your answer, being sure to talk about sample size, effect size, the results of the significance test, and practical considerations.

Choose the Correct Statistic

You now know three different statistics to test hypotheses about sample means: (1) the *z* for a sample mean, (2) the single-sample *t*, and (3) the related samples *t*. The goal of each test is described in the table Appendix J. Use this table to determine which statistic should be used in each of the following research scenarios.

48. Can people accurately judge how much time has passed? To answer this question, a researcher asks 35 students to sit in a room, alone, doing nothing. At the end of 8 minutes, she asks the students to estimate how long they were in the room. The average estimate was 10.1 minutes with a standard deviation of 2.34 minutes. Is 10.1 significantly different from 8?

49. To determine if yoga makes people more flexible, a sports psychologist recruits 25 runners to participate in a study. He first measures the flexibility of each runner by having runners sit with their legs extended in front of them and seeing how far they can stretch. The scores are recorded in centimeters and are normally distributed. The 25 runners attend yoga classes three times a week, and their flexibility is measured again. The mean score before yoga was 12.4 cm, and the mean score after yoga was 18.1 cm. Did flexibility increase significantly after the yoga class?

50. A recent nationwide study revealed that people eat an average of just 2.9 servings of fruits and vegetables per day. A nutritionist wonders if people who live in areas where fresh fruits and vegetables are easy to grow year-round (e.g., California) eat more fruits and vegetables than the national average. To test this hypothesis, the nutritionist asks 100 residents of California to record the average number of servings of fruits and vegetables they consumed in 1 week. The mean number of servings consumed by 100 residents in California was 3.3, with a standard deviation of 1.7. Does availability of fruits and vegetables affect fruit and vegetable consumption?

51. The same nutritionist wonders if gardening may encourage people to eat more vegetables. To test this, he recruits 23 people who have yards but do not currently have a garden. For 1 week prior to starting the garden, the participants record the number of servings of vegetables they consumed. After that week, the nutritionist works with a master gardener to teach the participants how to grow vegetables in their yards. Near the end of the growing season, the nutritionist again asks the participants to record the average number of servings of vegetables they consumed in 1 week. Did the program significantly increase the servings of vegetables eaten?

52. Intelligence tests scores are influenced by a number of factors, including nutrition. A health psychologist wonders if consumption of fruits and vegetables is associated with greater intelligence. Scores on a particular intelligence test are normally distributed, with a mean of $\mu = 100$ and a standard deviation of $\sigma = 15$. To determine if fruit and vegetable consumption is associated with greater intelligence, the psychologist recruits a sample of 43 people who eat at least seven servings of fruits and vegetables a day for a month and then measures their intelligence. The average intelligence test score for the sample was 104.32. Did the diet affect intelligence scores?

Activity 9.2: Combining Significance Testing, Effect Sizes, and Confidence Intervals

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Compute a confidence interval for a related mean difference by hand
- Interpret a confidence interval for a related mean difference
- Interpret SPSS output for confidence intervals
- Describe the distinct purposes of significance testing, effect sizes, and confidence intervals

Quick Review of CIs

In Activity 8.2, you learned how to combine a single-sample *t* test, an effect size, and two kinds of confidence intervals (CIs) to evaluate the job satisfaction of IM-ANX's highly trained employees. Specifically, you learned that significance tests, effect sizes, and CIs each provide researchers with distinct information about research results.

1. As a quick review, identify which statistical procedure provides researchers with each of the following types of information. Use "ST" for significance test, "ES" for effect size, and "CI" for confidence interval.

   _____ a. Provides a range of plausible values for a population parameter

   _____ b. Describes a treatment's impact in standard deviation units

   _____ c. Helps estimate a population mean (or mean difference) based on sample data

   _____ d. Helps determine if a study's result is likely to have occurred because of sampling error

   _____ e. Helps determine how well a treatment worked

   _____ f. Allows you to determine if a difference between means is statistically significant

2. In Activity 8.2, you learned that all CIs consist of a point estimate and a margin of error that is added to and subtracted from this point estimate. To compute a 95% CI for a population mean, the _____ is the point estimate.

   a. two-tailed $\alpha = .05$ critical *t* value

   b. *SEM*

   c. sample mean

   d. population mean

3. The margin of error for a 95% CI for a population is computed by multiplying which of the following two values together? (Choose two.)
   a. Two-tailed $\alpha = .05$ critical t value
   b. *SEM*
   c. Sample mean
   d. Population mean

4. To compute a 95% CI for a population *mean difference* when a study used only a single-sample, _____ is the point estimate.
   a. a two-tailed $\alpha = .05$ critical t value
   b. a *SEM*
   c. a mean difference observed between the sample mean and a population mean (or value of interest)

5. The margin of error for a 95% CI for a population *mean difference* when a study used a single sample is computed by multiplying which of the following two values together? (Choose two.)
   a. Two-tailed $\alpha = .05$ critical t value
   b. *SEM*
   c. Sample mean
   d. Population mean

Computing CIs for Related Samples

This activity has similar goals to that of Activity 8.2. You will integrate information from significance tests, effect sizes, and CIs into an APA-style research summary. However, in this activity, you will use a third type of confidence interval that estimates a population mean difference when a study uses two related samples (e.g., a pre–post or matched samples design). This confidence interval has the same three parts as all other confidence intervals: a point estimate, a critical t value, and an *SEM*.

Evaluating a Commercial's Effect on Voters (Pilot Study)

In hopes of improving Mrs. Puppet's image as a candidate for mayor, her campaign manager Mr. Strings created a commercial. Before airing the commercial, he hires your consulting firm to determine if the commercial would actually improve voters' attitudes toward Mrs. Puppet. As a first step, your firm conducts a very small pilot study to help those who will collect the data understand the experimental procedures. The firm wants to make sure the experimental procedures are well understood before it doles out all the cash needed for the large study. The firm obtains a sample of 10 undecided voters to participate in the study. First, each voter rated their current feelings toward Mrs. Puppet by answering several questions such as the following:

1	2	3	4	5	6	7	8	9
Definitely would not vote for her							Definitely would vote for her	

Then, all participants watched the commercial and answered all of the preceding questions a second time. Obviously, if the commercial works, approval ratings should be higher after voters view the commercial.

6. Even though this was just a pilot study, the firm is curious about the results. What statistical procedure should you use to determine if the change in voter ratings was likely to have occurred due to sampling error?
   a. Significant testing
   b. Effect size
   c. Confidence interval

7. Which of the following should you use to determine if the commercial worked? (This is the same question as the previous question, just worded differently.)
   a. *z* for a sample mean, one-tailed
   b. Single-sample *t* test, one-tailed
   c. Related samples *t* test, one-tailed
   d. Independent *t* test, one-tailed

8. For the purposes of doing hand calculations, you are going to work with a small sample size. Because of this small sample, which assumption is most likely to be violated in this study?
   a. Independence
   b. Appropriate measurement of the IV and the DV
   c. Normality
   d. The small sample size is likely cause all of the assumptions to be violated.

9. How large should the sample size be to address the violation identified in the previous question?
   a. At least 100
   b. At least 30

10. What are the one-tailed null and research hypotheses for this study? Put H_0 next to the null hypothesis and H_1 next to the research hypotheses.
    a. _____ The population of voters' mean rating of the mayor will be *higher after watching* the commercial than before watching the commercial.
    b. _____ The population of voters' mean rating of the major will be *lower after watching* the commercial than before watching the commercial.
    c. _____ The population of voters' mean rating of the mayor will *not be higher after watching* the commercial than before watching the commercial.
    d. _____ The population of voters' mean rating of the mayor will *not be lower after watching* the commercial than before watching the commercial.
    e. _____ The population of voters' mean rating of the mayor will be *different after watching* the commercial than before watching the commercial.
    f. _____ The population of voters' mean rating of the mayor will *not be different after watching* the commercial than before watching the commercial.

11. What is the *one-tailed* critical value for this study? _____

12. The data from the pilot study are listed below. Compute the significance test you chose as appropriate for this study. The mean of the previewing (M_{pre}) and postviewing (M_{post}) scores are provided below. You will need to compute the D for each participant, the mean of the difference scores (M_D), and the standard deviation of the difference scores (SD_D) and the obtained t value.

Previewing	*Postviewing*	*Difference Scores (D)*
4	5	
3	5	
3	4	
5	5	
4	6	
4	5	
6	7	
2	3	
4	4	
5	7	
$M_{pre} = 4.00$ $SD_{pre} = 1.15$	$M_{post} = 5.10$ $SD_{post} = 1.29$	$M_D =$ $SD_D =$

13. Should you reject the null hypothesis?
    a. Yes, the t value was less than .05.
    b. Yes, the t value was in the critical region.
    c. No, the t value was greater than .05.
    d. No, the t value was not in the critical region.

14. The significance test indicates that
    a. the increase in voters' ratings of Mrs. Puppet seem likely to have occurred by sampling error; the commercial did not work.
    b. the increase in voters' ratings of Ms. Puppet seem unlikely to have occurred by sampling error; the commercial did work.

15. Now compute an effect size for this pilot study.

16. The effect size you just computed was really large. It suggests that the commercial increased voters' ratings of Mrs. Puppet by 1.49 _____.
    a. sampling errors
    b. SEMs
    c. standard deviations
    d. population means

17. Next you want to determine how confident you can be in the amount of change this commercial may cause in the population of voters. To do this, you can compute a 95% CI around the *change* this commercial caused in the sample (i.e., +1.1). In this situation, the difference between the mean of previewing ratings and the mean of the postviewing ratings would be the _____ for the confidence interval.
    a. *t* score
    b. confidence level
    c. point estimate
    d. margin of error
    e. critical *t* value
    f. *SEM*

18. The margin of error that is added to and subtracted from the point estimate is computed by multiplying the correct *SEM* and critical *t* value for this situation. In this study, participants are measured twice, once before watching the commercial and then a second time after watching it. Therefore, the correct *SEM* would be the same equation used for a/an
    a. one-sample *t* test; $SEM_s = \frac{SD}{\sqrt{N}}$.
    b. independent *t* test; $SEM_i = \sqrt{\frac{SD_p^2}{n_1} + \frac{SD_p^2}{n_2}}$.
    c. related samples *t* test; $SEM_r = \frac{SD_D}{\sqrt{N}}$.

19. The other value contributing to the margin of error is the critical *t* value for a 95% CI, which is always equal to the *two-tailed* .05 alpha value from Appendix B. In this study, the number of participants was $N = 10$, so the critical *t* value would be _____. (Important note: Even though the significance test was one-tailed, you *always* need the two-tailed .05 value to compute a 95% CI.)

20. So, when you are computing a 95% CI, the t_{CI} will always be the value from the _____ even if the significance test was a one-tailed test.
    a. two-tailed .05 critical table
    b. one-tailed .05 critical table
    c. two-tailed .01 critical table
    d. one-tailed .01 critical table

21. Compute the 95% confidence interval for the *change* in voters' ratings resulting from the commercial.

22. What does the point estimate represent?
    a. It is an estimate of the amount of attitude change in the population.
    b. It is the actual amount of attitude change in the population.
    c. It is the estimate of how much the sample liked Mrs. Puppet.
    d. It is the actual amount the population liked Mrs. Puppet.

23. What do the lower and upper bounds represent?
    a. The lower and upper limits contain the range of plausible values for the amount of attitude change to expect in the population.
    b. The lower and upper limits for the amount of attitude change in the sample.
    c. The range of how much the sample liked Mrs. Puppet.
    d. The range of the actual amount the population liked Mrs. Puppet.

24. To complete the summary, you will first need to compute two CIs. You will need the CI around the mean approval ratings before watching the commercial and also the CI for the ratings after watching it. Compute both in the space provided. Refer to Activity 8.2 for guidance in computing these confidence intervals if necessary.

25. Someone else wrote up a draft of your firm's final report and you need to proofread it and make final additions before the boss gets the report.

    The results of the pilot study indicated that approval ratings were significantly higher after watching the commercial ($M = 5.10$, $SD = 1.29$), 95% CI [_____, _____], than before watching the commercial ($M = 4.00$, $SD = 1.15$), 95% CI [_____, _____], $t($ _____) = _____, $p =$ _____, $d =$ _____, 95% CI [_____, _____]. These results suggest that the commercial may be effective because the change in attitude was statistically significant and we can be _____% confident that the true value of the attitude change is between _____ and _____. This is a large effect. Although these results are promising, we should be cautious in interpreting them because the sample size is too _____ to draw firm conclusions.

26. Now that all the "kinks" have been worked out in the experimental procedure, your firm is ready to run the larger study to test the campaign commercial. The null and research hypotheses will be the same as the pilot study, but the critical value will be different. Why?
    a. The alpha level will be larger.
    b. The sample size will be larger.
    c. The observed t value will be larger.

27. Your team collected all the data and entered it into SPSS. Now you have to run the analyses. Find the "MayorCommercial.sav" data file and do the following:

    - Click on Analyze, Descriptive Statistics, and then Explore.
    - Move the variable(s) of interest (in this case PreTestRating and PostTestRating) into the Dependent List box.
    - Click on the Statistics button.
    - In the Explore:Statistics box, select Descriptives and make sure the Confidence Interval for Mean is set at 95%.
    - Click on the Continue button and then on the OK button to run the analysis.

    Based on the SPSS output, how many people participated in this study?

    $N =$ _____

28. What were the mean approval ratings on the pretest and the posttest?

    Pretest mean= _____ Posttest = mean_____

29. Of course, the above previewing and postviewing means are from a sample. The campaign really wants to know what impact the commercial would have on the entire *population* of undecided voters. The possibility of sampling error implies that the actual population parameters for previewing and postviewing ratings may vary considerably. What does the 95% confidence interval around the *previewing* mean?

    Previewing LB = _____ Previewing UB = _____

30. What does the 95% confidence interval around the *postviewing* mean?

    Postviewing LB = _____ Postviewing UB = _____

31. Now you need to determine if the commercial changed voters' ratings or if the pre–post ratings difference was likely to have been created by sampling error. Which statistical procedure addresses this question?

    a. Significance test
    b. Effect size
    c. Confidence interval

32. Which of the following would enable you to determine if the commercial increased voters' ratings?

    a. Single-sample *t* test
    b. Related samples *t* test
    c. Independent samples *t* test

33. The campaign is only interested in using the commercial if it actually *increases* Mrs. Puppet's approval ratings. Should your firm use a one-tailed or two-tailed significance test?

    a. One-tailed
    b. Two-tailed

34. What are the one-tailed null and research hypotheses for this study? Write H_0 next to the null hypothesis and H_1 next to the research hypothesis for this *t* test.

    a. _____ The population of voters' approval ratings will *not be higher after watching* the commercial ($\mu_D < 0$).

    b. _____ The population of voters' approval ratings will *be higher after watching* the commercial ($\mu_D > 0$).

    c. _____ The population of voters' approval ratings will *be different after watching* the commercial ($\mu_D \neq 0$).

    d. _____ The population of voters' approval ratings will *not be different after watching* the commercial ($\mu_D = 0$).

    Now that you know what significance test to run, you need to run it. Run the related samples *t* test by doing the following:

    - Click on Analyze, Compare Means, and then Paired Samples *t* test.
    - Move the variable(s) of interest (in this case, PostTestRating and PreTestRating) into the Paired Variables box. (Note: The output will be easier to understand if you move PostTestRating into the Variable 1 box and PreTestRating into the Variable 2 box.)
    - Click OK.

35. What is the *t* value and *p* value for the larger commercial study?

    $t(32) =$ _____, $p =$ _____

36. Choose the best interpretation of this *p* value. (Choose all that apply.)

    a. The probability of obtaining a sample mean difference of .58 or more extreme if the null hypothesis is true.

    b. The probability of rejecting the null hypothesis.

    c. In this study, a mean difference of .58 would occur less than 1 time out of 1,000 if the null were true.

37. Should you reject or fail to reject the null hypothesis?

    a. Reject, $p < .05$

    b. Fail to reject, $p > .05$

38. Which of the following statements is the best verbal summary of the results of the significance test?

    a. The approval ratings were significantly *different after watching* the commercial than before watching the commercial.

    b. The approval ratings were *not* significantly *different after watching* the commercial than before watching the commercial.

    c. The approval ratings were significantly *higher after watching* the commercial than before watching the commercial.

    d. The approval ratings were significantly *lower after watching* the commercial than before watching the commercial.

39. The significance test suggests that the commercial worked. But, how well did it work? To answer this question, you need to use which of the following statistical procedures?
    a. A significance test
    b. An effect size
    c. A confidence interval

40. SPSS does not compute *d* for a related samples *t* test. Pull the values you need from the SPSS output and compute it by hand. What is the value for *d*?

41. The sign or direction of an effect size is completely dependent on whether the numerator of *d* is computed as (Pre – Post) or (Post – Pre). First, determine which way SPSS computed the mean difference and then choose a correct interpretation of the computed *d*.
    a. Approval ratings were 1.03 of a standard deviation lower after watching the commercial.
    b. Approval ratings were 1.03 of a standard deviation higher after watching the commercial.
    c. The researcher failed to reject the null hypothesis.
    d. The researcher rejected the null hypothesis.

42. Choose the best interpretation of the size of this effect.
    a. Small
    b. Small to medium
    c. Medium
    d. Medium to large
    e. Large

43. Is this effect size good news or bad news for the campaign? Explain your answer.
    a. Bad news! Although the effect size was large, the result was not statistically significant, suggesting that the results might be due to sampling error.
    b. Bad news! The effect size was large, but the results were in the wrong direction, suggesting that the commercial had a negative effect on ratings.
    c. Good news! The effect size was large and the results was also statistically significant.

44. What is the mean difference on your SPSS output? Mean difference = _____

45. This exact mean difference is based on samples, so the mean difference that would actually occur in the population might vary from this value considerably. What is the range of plausible values for *approval rating change* the campaign could reasonably expect if the *population* viewed the commercial? In other words, what is the 95% CI around this *mean difference*?

    LB = _____ UB = _____

46. Now that the large study is complete, your boss wants a report. Use the report you generated for the pilot study as a model for this new report on the large study. When reporting numbers in APA style, round to the second decimal place. Make sure you provide a final recommendation to Mrs. Puppet's campaign. Based on these data, what would you say to Mrs. Puppet's campaign

about the commercial? (Be sure to provide an interpretation of the significance test, effect size, and the most important 95% CI.)

The results of the study indicated that approval ratings were significantly higher after watching the commercial (M = _____, SD = _____), 95% CI [_____, _____] than before watching the commercial (M = _____, SD = _____), 95% CI [_____, _____], t(_____) = _____, p _____, d = _____, 95% CI [_____, _____]. These results suggest that the increase in approval ratings is statistically significant, and we are _____% confident that mean amount of attitude increase in the population is likely to be between _____ and _____. This is a _____ effect size. A larger sample size would still be preferable, but we are much more confident in these results than the results of the pilot study.

47. One person on your team notices that the obtained t value in the pilot study was 4.71 and the obtained t value in the larger study was 5.90. Do you think it is a good idea to compare t values from studies with different sample sizes? (Hint: Is the obtained t value affected by sample size?)
    a. Yes, the study with the larger t value has stronger evidence that the commercial worked.
    b. No, t values are highly influenced by sample size, so comparing t values across studies with difference sample sizes is unfair.

48. One person on your team notices that the p value in the pilot study was .001 (two-tailed), and the p value in the larger study was less than .001 (two-tailed). Do you think it is a good idea to compare p values from studies with different sample sizes? (Hint: Is the p value affected by sample size?)
    a. Yes, the study with the smaller p value has stronger evidence that the commercial worked.
    b. No, p values are highly influenced by sample size, so comparing p values across studies with difference sample sizes is unfair.

49. One person on your team notices that the effect size in the pilot study was d = 1.49 and the effect size in the larger study was d = 1.03. Do you think it is a good idea to compare d values from studies with different sample sizes? (Hint: Is the d affected by sample size?)
    a. Yes, the purpose of effect size is to compare the effects across studies. Even when the sample sizes differ.
    b. No, effect sizes should only be compared when the sample sizes are equal across studies.

50. Effect sizes are designed to help researchers compare results across studies. If studies are on the same topic, use the same experimental procedures, and yet produce very different effect sizes, it should be a "red flag" to researchers. Researchers should try to explain the very different results. What might explain the very different effect sizes generated by the pilot study and the larger study? (Choose all that apply.)
    a. Greater chance of sampling error in the pilot study than in the larger study.
    b. If the experimental procedure was "messier" in the pilot study, which is often the case, the effect size for the pilot study could be misleading.

51. Which study has the narrower (more precise) confidence interval?
    a. Study 1 because the sample size is much smaller
    b. Study 2 because the sample size is much larger

Chapter 9 Practice Test

1. Researchers have long known that people feel lonely when they are ostracized by other people. A researcher wonders if being ostracized by a computer also makes people feel lonely. To test this hypothesis, the researcher recruits 18 people to participate in a study and measures their current feelings of loneliness using the Social Loneliness Scale (SLS). This scale yields scores between 0 and 11, with higher scores meaning greater feelings of loneliness. The SLS scores form an interval scale that is normally distributed. After completing the pretest measure of loneliness, the participants play a game called Cyberball. In the game, participants simply play catch with computer-generated players. The participants in the study know that the players in the game are computer controlled and are not controlled by human players. When participants begin the game, the computer-generated players pass the ball back and forth with each other and with the participant. After 30 throws, the computer-generated players stop passing the ball to the participant. This exclusion (i.e., ostracism) continues for 20 passes before the games end. After playing the game, all participants complete the SLS again. What is the one-tailed research hypothesis for this study?

   a. Participants will be less lonely after playing the game than before playing the game.
   b. Participants will be more lonely after playing the game than before playing the game.
   c. Participants will not be less lonely after playing the game than before playing the game.
   d. Participants will not be more lonely after playing the game than before playing the game.

2. What is the one-tailed null hypothesis for this study?

   a. Participants will be less lonely after playing the game than before playing the game.
   b. Participants will be more lonely after playing the game than before playing the game.
   c. Participants will not be less lonely after playing the game than before playing the game.
   d. Participants will not be more lonely after playing the game than before playing the game.

3. Match the assumption to the fact that suggests that assumption was met.

   ___ Independence
   ___ Appropriate measurement of the IV and the DV
   ___ Normality

   a. The SLS scores are normally distributed.
   b. Each participant plays the game with no one else around.
   c. The pregame and postgame conditions are clearly defined. Responses to the questionnaire are measured on an interval scale.

4. Compute the degrees of freedom for this *t* test.

   a. 18
   b. 9
   c. 17
   d. 36

Pretest	Posttest
4	4
5	5
6	7
8	6
4	5
1	3
2	3
6	5
5	5
3	3
4	5
2	3
2	4
4	4
5	6
4	5
3	4
6	8

5. What is the critical value for this *t* test (use an alpha of .05)?
   a. 1.7396
   b. 1.7341
   c. 2.1098
   d. 2.1009

6. Compute the *t* statistic for these data. You may use your calculator, not SPSS.
   a. $t = -1.04$ or 1.04
   b. $t = -2.50$ or 2.50
   c. $t = -.61$ or .61
   d. $t = -1.52$ or 1.52

7. Should you reject or fail to reject the null hypothesis?
   a. Reject
   b. Fail to reject

8. Compute the effect size (*d*).
   a. .61
   b. 2.50
   c. 1.04
   d. .59

9. How is this effect best described?
   a. Small
   b. Small to medium
   c. Medium
   d. Medium to large
   e. Large

10. Fill in the blanks.

    Participants reported greater loneliness after being ostracized by the computer-generated players ($M =$ _____, $SD =$ _____) than before being ostracized ($M =$ _____, $SD =$ _____). This difference of _____ was statistically significant, $t($_____$) =$ _____, $p <$ .05, $d =$ _____.

11. How quickly do we decide if we like someone or not? To answer this question, a researcher creates a video of a 20-year-old male talking about his childhood. All participants watch the video without sound and are told that they should try to form an impression of this person based on his body language. All participants watch the first 10 seconds of the video and then rate how much they like this person on a scale from 1 to 10, where 1 = *not at all* and 10 = *a lot*. Then, they continue to watch the video for 5 more minutes. At the end of the 5 minutes, the participants again rate how much they like this person. Do a two-tailed test to determine if the ratings changed between the first and second ratings. What is the null hypothesis for this study?

a. $\mu_D = 0$
b. $\mu_D \neq 0$
c. $\mu_D \geq 0$
d. $\mu_D \leq 0$

12. What is the research hypothesis for this study?

a. $\mu_D = 0$
b. $\mu_D \neq 0$
c. $\mu_D \geq 0$
d. $\mu_D \leq 0$

13. The output below is from the related samples *t*.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	FirstRating	5.3750	8	1.30247	.46049
	SecondRating	5.8750	8	1.55265	.54894

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	FirstRating & SecondRating	8	.733	.039

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	FirstRating - SecondRating	-.50000	1.06904	.37796	-1.39374	.39374	-1.323	7	.227

How many people participated in the study?

a. 7
b. 8
c. 14
d. 16

14. What was the average difference between the first and second ratings?

a. 5.375
b. 5.875
c. −.50
d. 1.232
e. .37796
f. .47

15. What difference would you expect between the first and second ratings if the null hypothesis is true?
    a. 0
    b. .50
    c. 1.65
    d. .37796

16. Should the researcher reject or fail to reject the null hypothesis ($\alpha = .05$)?
    a. Reject
    b. Fail to reject

17. Compute the effect size (d) for this study.
    a. −5.375
    b. −5.875
    c. −.50
    d. −1.232
    e. −.37796
    f. −.47

18. SPSS computes the 95% confidence interval around the mean difference. What is the range of plausible values for the mean difference?
    a. 5.37, 5.88
    b. 1.30, 1.55
    c. −1.39, .39

19. If you were to compute the 95% confidence interval around the mean difference, what value would you use for the SEM_r?
    a. .378
    b. .460
    c. .549

20. If you were to compute the 95% confidence interval around the mean difference, what value would you use for the t_{CI}?
    a. 2.3646
    b. 1.8946
    c. 1.8595
    d. 2.3060

21. Choose the best APA-style summary of these results.
    a. The difference between the ratings after watching the video for 10 seconds ($M = 5.38$, $SD = 1.30$) and after watching the video for 5 minutes ($M = 5.88$, $SD = 1.55$) was not statistically significant, $t(7) = -1.32$, $p = .23$, $d = .47$, 95% CI [−1.39, .39].

b. The difference between the ratings after watching the video for 10 seconds ($M = 5.38$, $SD = 1.30$) and after watching the video for 5 minutes ($M = 5.88$, $SD = 1.55$) was statistically significant, $t(7) = -1.32$, $p = .23$, $d = .47$, 95% CI [−1.39, .39]. Scores were higher after 5 minutes than after 10 seconds.

c. The difference between the ratings after watching the video for 10 seconds ($M = 5.38$, $SD = 1.30$) and after watching the video for 5 minutes ($M = 5.88$, $SD = 1.55$) was not statistically significant, $t(7) = -1.32$, $p = .23$, $d = .47$, 95% CI [−1.39, .39]. Scores were significantly higher after 5 minutes than after 10 seconds.

d. The difference between the ratings after watching the video for 10 seconds ($M = 5.38$, $SD = 1.30$) and after watching the video for 5 minutes ($M = 5.88$, $SD = 1.55$) was statistically significant, $t(7) = -1.32$, $p = .23$, $d = .47$, 95% CI [−1.39, .39].

22. What recommendation would you give this researcher?

a. Redo the study with a larger sample size

b. Redo the study with a larger population

c. Redo the study with an alpha of .01.

23. Which of the following is the best explanation for what it means to reject the null hypothesis?

a. You are unlikely to obtain a sample mean difference this far from the population mean difference simply because of sampling error, and thus, it is likely that the difference is due to the independent variable.

b. There is only a 95% chance that the research hypothesis is true.

c. The distribution of sample means for the research hypothesis is unlikely to occur if the null hypothesis is true.

24. Researchers only compute effect sizes after they reject the null hypothesis.

a. True

b. False

25. What does it mean when researchers say that a result is statistically significant?

a. The effect size is medium or large (i.e., not small).

b. They rejected the null hypothesis.

c. The result is practically important.

26. When do you reject the null hypothesis?

a. When the *p* value is less than the alpha

b. When the *p* value is greater than the alpha

27. When do you reject the null hypothesis?

a. When the *t* value is in the critical region

b. When the *t* value is outside of the critical region

28. A researcher conducts a *t* test and concludes that the treatment worked. What type of error might this researcher have made?

a. Type I

b. Type II

29. As sample size increases, which of the following also tends to increase?
    a. Type I error rate
    b. The size of the critical region
    c. Statistical power
    d. All of the above

30. As the obtained t value increases, what happens to the p value?
    a. Increases
    b. Decreases
    c. The relationship between the t and p values is not predictable.

REFERENCE

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge Academic.

CHAPTER 10

Independent Samples *t* Test

LEARNING OBJECTIVES

After reading this chapter, you should be able to do the following:

- Explain when to use an independent samples *t* test
- Explain the logic of the independent samples *t* test
- Write null and research hypotheses using symbols and words for both one- and two-tailed tests
- Compute degrees of freedom and define a critical region for both one- and two-tailed tests
- Compute an independent samples *t* using a calculator and SPSS
- Compute and interpret an effect size (*d*)
- Summarize the results of the analysis using American Psychological Association (APA) style

INDEPENDENT SAMPLES *t*

Thus far, you have learned to use two different *t* tests: the single-sample *t* and the related samples *t*. You should use the single-sample *t* whenever you want to compare a sample mean to a population mean (or a value of theoretical interest) but you do not know the population's standard deviation. You should use the related samples *t* test when comparing the sample means from either the same individuals measured at two different times or pairs of matched people measured under different conditions. When you need to compare two sample means that are *unrelated,* you must use a third *t* test: an independent (samples) *t* test.

An example may help illustrate the difference between these three *t* tests. Suppose that you are interested in the effect of a drug on weight loss. To test the efficacy of this drug, you could give the drug to a sample of people and, after they are on the drug for 1 month, compare the mean pounds lost by the sample to zero pounds. In this case, you would use a single-sample *t*. However, if you measured the sample's mean weight before taking the drug and then again a month later after taking the drug, you would use a related samples *t*. Finally, you could also give one sample of people the drug and another sample a placebo. After both samples had taken their respective "drugs" for 1 month, you could compare the mean weight loss of the two samples. In this final option, you would use an independent samples *t* test because the two samples contain different and unmatched people.

The independent *t* test uses two samples from the population to represent two different conditions. As in the example earlier, it is often the case that one sample is intended to represent what the population would be like if nothing were done to it (i.e., a control condition), and another sample is intended to represent what the population would be like if it were given some treatment (i.e., experimental condition). The objective of the independent samples *t* test is to determine if the difference between the two sample means is likely or unlikely to be due to sampling error.

The logic of the independent *t* test is similar to that of the *z* for the sample mean, the single-sample *t* test, and the related samples *t* test. All four tests compute a ratio, specifically, the observed deviation between two means over the deviation expected due to sampling error:

$$\text{Obtained } t \text{ or } z = \frac{\text{Observed ddifference between the means}}{\text{Mean difference expected due to sampling error}}.$$

For all of these tests, if the null hypothesis is true, the obtained *t* or *z* should be zero. However, if the null hypothesis is false, the obtained *t* or *z* should be far from zero.

Reading Question

1. Which significance test should you use to determine if the difference between two unrelated *samples is likely to be due to sampling error?*
   a. *z* for a sample mean
   b. Single-sample *t* test
   c. Independent *t* test

Reading Question

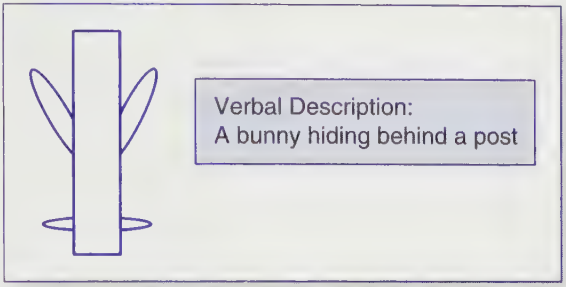
2. If the null hypothesis is true, the *z* for a sample mean, single-sample *t* test, related samples *t* test, and the independent *t* test all expect an obtained value close to
   a. +1.65.
   b. +1.96.
   c. +1.00.
   d. 0.

The independent *t* test compares two sample means from two unrelated (i.e., independent) samples/groups. For example, suppose you and your friend Bill team up on a research project for your human cognition course. The two of you want to test how verbal labels influence participants' memory of pictures. You investigate this question by showing all participants a series of 25 simple line drawings like those shown in Figure 10.1 and asking them to recall the drawings 10 minutes later. However, half of the participants only saw the drawings while the other half *also* saw a verbal description of each drawing similar to that in Figure 10.1. Your study is similar to one conducted by Bower, Karlin, and Dueck (1975).

You need to compare the memory of those who saw the drawings *and* verbal descriptions (i.e., the experimental group) to the memory of those who only saw the drawings (i.e., the control group). In this situation, you created two samples and gave each a different treatment. You then measured the mean number of drawings each sample recalled. The sample mean from the control group estimates what the population's mean memory score would be if the population only saw the drawings with no verbal labels. In contrast, the other sample estimates what the population's mean memory score would be if everyone saw the drawings *and* verbal labels. You can use an independent samples *t* test

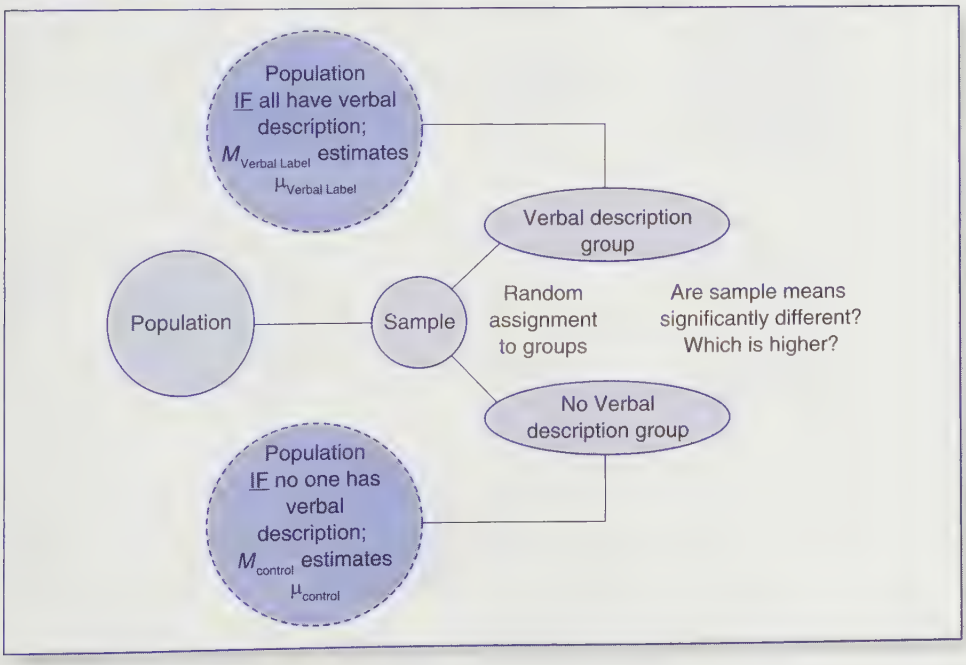
to determine if the difference between these two sample means was likely or unlikely to have occurred due to sampling error. If the experimental group had a higher mean *and* the obtained *t* value is in the critical region, you could conclude that the verbal descriptions increased memory scores. If, however, the verbal description group had a significantly lower mean, you could conclude that the verbal descriptions decreased memory scores. Figure 10.2 illustrates this research scenario.

Figure 10.1 Simple Line Drawing and Its Verbal Description Used in Your Human Memory Experiment



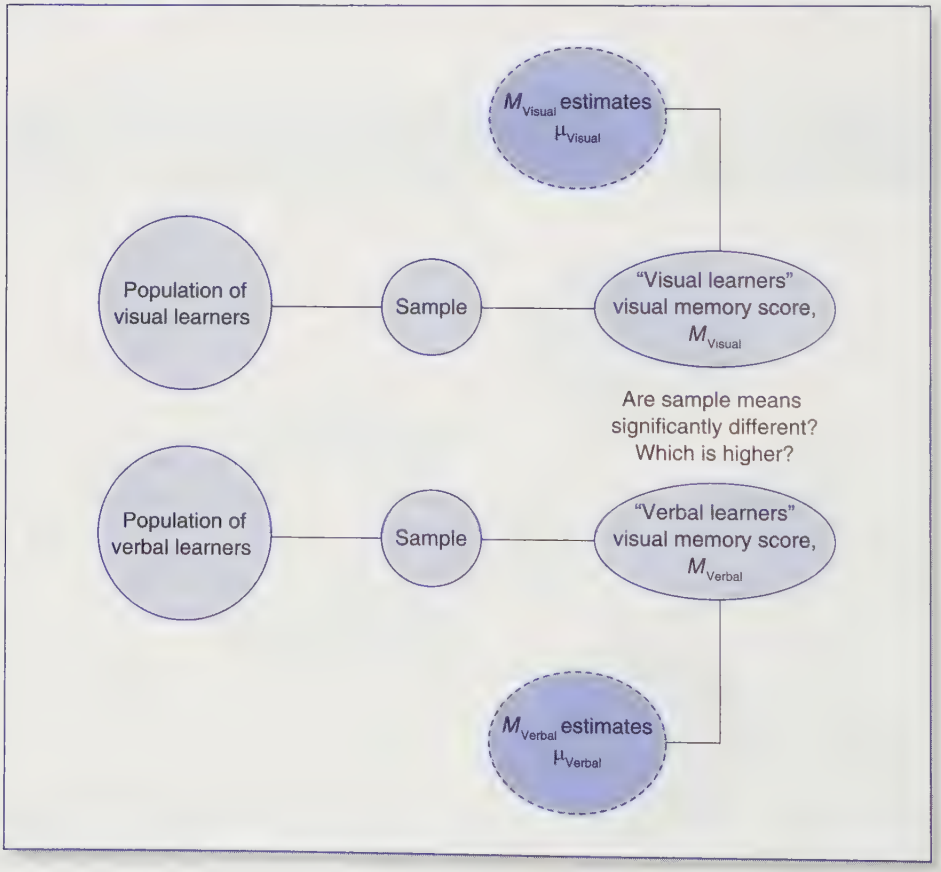
In your study, you took one sample from a single population and then divided that sample to create two different groups; one received no verbal labels (i.e., control condition) and the other received verbal labels (i.e., experimental condition). You essentially took people who were similar and made them different by giving them different treatments (i.e., you gave them different levels of the IV [independent variable]). You then used the independent *t* test to determine if the different IV levels affected memory differently.

Figure 10.2 Research Approach of Creating Two Different Samples to Infer What the Population Would Be Like Under Two Different Situations



An independent *t* test can also be used to compare two distinct populations of people who are already different in some way. In other words, an independent *t* test can compare groups with a pre-existing difference. For example, suppose you and Bill decide to design another memory experiment to test a different research question. You want to test a theory of learning called "learning styles." The theory proposes that some people are "visual learners" and others are "verbal learners." According to learning styles theory, "visual learners" should learn *visual* material better than "verbal learners" learn *visual* material. You want to test this prediction by determining if the mean visual memory score of "visual learners" is significantly higher than that of "verbal learners." To test this hypothesis, you take a sample of "visual learners" and a sample of "verbal learners," show both groups 25 simple line drawings (i.e., visual information), and measure their ability to recall the drawings 2 days later by asking them to re-create the 25 drawings. You then use the independent *t* test to determine if the two samples' mean visual memory scores are significantly different. Figure 10.3 illustrates the "visual learner" versus "verbal learner" scenario.

Figure 10.3 Research Approach of Obtaining Two Different Samples to Infer Difference Between Two Different Populations



Reading Question

3. An independent *t* test can be used to compare differences between people that
   a. are created by the researcher by providing different IV levels.
   b. already exist in different populations of people.
   c. Both of the above

CONCEPTUAL FORMULA FOR THE INDEPENDENT SAMPLES *t*

Regardless of whether you have means from two groups with a preexisting difference or from two groups in which you created a difference, the same independent *t* test formula is used. The independent *t* test is the ratio of the observed mean difference over the difference expected due to sampling error:

$$t = \frac{\text{Samples' mean difference} - \text{Populations' mean difference expected if } H_0 \text{ is true}}{\text{Mean difference expected due to sampling error}}.$$

As indicated in the earlier "logical formula," there are three terms in this *t* test. The two samples' mean difference is determined by the data. In the "visual learners" versus "verbal learners" scenario, this term is the actual difference between the memory scores of "visual" and "verbal" learners. The null hypothesis determines the other term in the numerator. For example, if learning style (visual vs. verbal) has *no* impact on memory scores, we would expect the population mean memory score of the two groups to be the same and their mean difference to be zero. In most research situations (all situations in this book), the populations' mean difference that is expected if the null is true is zero. In other words, the numerator is simply the difference between the two sample means. The term in the denominator represents the amount of sampling error expected. In the following formula, M_1 and M_2 are the sample means of Group 1 and Group 2, respectively.

$$t = \frac{(M_1 - M_2)}{SEM_1}.$$

Reading Question

4. The independent *t* test is a ratio of the difference between two sample means over an estimate of
   a. sampling error.
   b. the standard deviation of the scores.
   c. variability.

Reading Question

5. The numerator of the independent samples *t* test is the difference between
   a. two sample means.
   b. a sample mean and a population mean.
   c. a sample mean and the null hypothesis.

TWO-TAILED INDEPENDENT *t* TEST EXAMPLE

After designing your study investigating the effect of verbal descriptions on memory for simple line drawings, you and your friend Bill discuss your predictions. Bill thinks that providing the verbal descriptions along with the line drawings will distract the participants from attending to the line drawings. So, Bill thinks that the mean memory score for the verbal description group will be *lower* than the mean memory score for the no verbal description group. You disagree. You think that the verbal descriptions will help give meaning to the otherwise abstract line drawings, and this greater meaning should increase memory scores. So, you think the verbal description group will have the *higher* mean memory score. Because there are two competing theories that make opposing predictions, you wisely choose to conduct a *two-tailed t* test. Twelve students volunteer to participate in your study. Half of them see the 25 line drawings *with* verbal descriptions and the other half see the 25 line drawings *without* verbal descriptions. You record each person's memory score. Memory scores form a normal distribution. You use a two-tailed *t* test with an alpha of .05 to test your research question.

Group 1: With Verbal Descriptions Group:

21, 22, 20, 20, 18, 20

Group 2: Without Verbal Descriptions Group:

19, 20, 19, 18, 16, 20

Step 1: Examine the Statistical Assumptions

You collected your data carefully, thereby satisfying the *data independence assumption*. The DV, number of line drawings correctly recalled, is on an interval/ratio scale, and the IV, presence or absence of verbal descriptions, identifies two different groups/conditions. Therefore, the study meets the *appropriate measurement of variables assumption*. When it comes to the *normality assumption*, distributions of memory scores tend to be normally shaped and therefore this assumption is likely met. This study will probably have very low statistical power and a lot of sampling error because of its very small sample sizes, and you will need to be very cautious interpreting the results. But you and Bill decide to analyze the data anyway.

The last assumption to consider is the *homogeneity of variances assumption*. For the independent *t* test, this assumption is that the two groups have similar variability in their memory scores. In previous chapters, we used the general rule that if the standard deviation in one condition is double that of another condition, this assumption *might* be violated. This is still a good guide to follow for the independent *t* test, but for the independent *t* test, there is a more precise way of assessing this assumption. As you know from the previous section, the obtained *t* value of an independent *t* test is the mean difference between the two conditions divided by expected sampling error. There are actually two ways to compute expected sampling error when doing an independent *t* test. One way assumes homogeneity of variance (i.e., that the two conditions have similar variances) and the other way does not. Which way is best depends on the data. If the variances are in fact similar, then computing sampling error by assuming equal variances is best. If, however, the variances are in fact very different, computing sampling error without assuming equal variances is best. In most cases, assuming equal variances will be the best, so that is what this book teaches you to do when doing hand computations. However, there are times when assuming unequal variances is better. Obviously, the amount of sampling error affects the obtained *t* value and your decision about rejecting the null.

Fortunately, SPSS computes both obtained *t* values automatically. Therefore, when using SPSS, you should know how to determine which obtained *t* value is the best for your data. SPSS provides a **Levene's test** to help you make this decision. We will describe this test in detail in the SPSS section of this chapter, but the main idea is that this test compares the variability in the two conditions, and the results of this test indicate if the two variances are similar enough to satisfy the equal variances assumption. *Based on the results of this Levene's test, you will choose between the* t *test that assumes equal variance or the one that does not.* Again, you will see this test and how to interpret its results in the SPSS section of the chapter. The results of Levene's test are generally consistent with the double standard deviation guideline you learned earlier in this text. After assessing the assumptions, you are ready to move on to the next step.

Reading Question

6. You use an independent samples *t* statistic when
   a. the IV defines two independent samples and the DV is measured on an interval/ratio scale.
   b. the IV defines two matched samples and the DV is measured on an interval/ratio scale.
   c. the IV defines one sample, the DV is measured on an interval/ratio scale, and the DV is measured twice on that same sample.
   d. the IV defines one sample and the DV is measured on an interval/ratio scale, and you do not know the population standard deviation.
   e. the IV defines one sample and the DV is measured on an interval/ratio scale, and you do know the population standard deviation.

Reading Question

7. Levene's test will help you determine
   a. whether or not the two sample means are significantly different from each other.
   b. if you should reject the null or fail to reject the null.
   c. if the two conditions have similar variances or variances that are very different (i.e., if the homogeneity of variance assumption is met or violated).
   d. if the sample size is sufficiently large.

Step 2: State the Null and Research Hypotheses Symbolically and Verbally

As mentioned above, you and Bill have opposing predictions about the effects of verbal labels on memory, so you correctly agree to use a two-tailed test. As shown in Table 10.1, the two-tailed research hypothesis states that the two means are different (i.e., are not equal). The null hypothesis is the exact opposite, stating that the two means are not different (i.e., are equal).

Reading Question

8. Which of the following represents the null hypothesis when doing a two-tailed independent test?
   a. $\mu_1 = \mu_2$
   b. $\mu_1 \neq \mu_2$

Table 10.1 Symbolic and Verbal Representations of Two-Tailed Research and Null Hypotheses for an Independent *t* Test

	Symbolic	*Verbal*	*Mean Difference Created by*
Research hypothesis (H_1)	H_1: $\mu_1 \neq \mu_2$ OR H_1: $\mu_1 - \mu_2 \neq 0$	If a population were given the verbal descriptions, their mean memory score *would be significantly different* from a population not given the verbal descriptions.	The verbal label affecting memory scores
Null hypothesis (H_0)	H_0: $\mu_1 = \mu_2$ OR H_0: $\mu_1 - \mu_2 = 0$	If a population were given the verbal descriptions, their mean memory score *would not be significantly different* from a population not given the verbal descriptions.	Sampling error

Reading Question

9. Which of the following is the best summary of the two-tailed research hypothesis?
   a. People in the verbal description group will have higher memory scores than people in the no verbal description group.
   b. The people in the verbal description group will have different memory scores than people in the no verbal description group.

Step 3: Compute the Degrees of Freedom and Define the Critical Region

When you computed the degrees of freedom (*df*) for the single-sample *t* test, the *df* formula was $df = (N - 1)$. The *df* formula for an independent *t* test is different because the independent *t* test uses two samples rather than just one sample. Therefore, you have to compute the *df* for each sample and combine them to get the *df* for the independent *t* test. The independent *t* test *df* formula is $df = (n_1 - 1) + (n_2 - 1)$, where n_1 and n_2 represent the sample sizes of the two samples, respectively.

In this case, the *df* is as follows:

$$df = (n_1 - 1) + (n_2 - 1) = (6 - 1) + (6 - 1) = 10.$$

You then use the correct *t* table of critical values and the $\alpha = .05$ criterion to find the critical value of *t* for this two-tailed independent *t* test. In this case, the critical value is 2.2281. This means that the two critical regions are $t \geq +2.2281$ and $t \leq -2.2281$.

Step 4: Compute the Test Statistic

4a. Compute the Deviation Between the Two Sample Means

As mentioned above, there are two terms in the numerator of the *t* statistic. The first $(M_1 - M_2)$ is the difference between the two sample means. The second $(\mu_1 - \mu_2)$ is the difference between the two

population means, assuming the null hypothesis is true. Although it is possible to test null hypotheses that predict a specific difference other than zero (e.g., $\mu_1 - \mu_2 = 10$), these types of tests are very rare and will not be covered in this text. For our purposes, $\mu_1 - \mu_2$ will always equal 0. Thus, the numerator is simply the difference observed between the two sample means:

$$(M_1 - M_2) = (20.17 - 18.67) = 1.50.$$

Reading Question

10. When computing the numerator of the independent samples *t* test, the population mean difference (i.e., $\mu_1 - \mu_2$) will always be
    a. 0.
    b. 10.
    c. the same as the sample mean difference.

4b. Compute the Expected Sampling Error

Computing the denominator, or expected sampling error, requires several steps. First, compute the standard deviation for Group 1 (SD_1), and then compute the standard deviation for Group 2 (SD_2). In this problem, the standard deviation for *each group of scores* is not given to you, so you must compute each of them. In other problems, this information may be provided. As you may recall, the computational formula for sum of squares is $SS = \sum X^2 - \frac{(\sum X)^2}{n}$. Sum all of the scores from one of the groups to find $\sum X$ for *that group of scores*. Then square every score and sum all of the squared scores to find $\sum X^2$ for *that group of scores*. In this example, the $\sum X$ for Group 1 (i.e., $\sum X_1$) = 121, and the $\sum X^2$ for Group 1 = 2,449. There were six scores in Group 1, so $n_1 = 6$. Therefore, the SS_1 and SD_1 are as follows:

$$SS_1 = \sum X^2 - \frac{(\sum X)^2}{n} = 2{,}449 - \frac{(121)^2}{6} = 2{,}449 - 2{,}440.17 = 8.83.$$

$$SD_1 = \sqrt{\frac{SS_1}{n-1}} = \sqrt{\frac{8.83}{5}} = 1.33.$$

The $\sum X$ for Group 2 (i.e., $\sum X_2$) = 110, and the $\sum X^2$ for Group 2 = 2,026. There were six scores in Group 2, so $n_2 = 6$. Therefore, the SS_2 and SD_2 are as follows:

$$SS_2 = \sum X^2 - \frac{(\sum X)^2}{n} = 2{,}102 - \frac{(112)^2}{6} = 2{,}102 - 2{,}090.67 = 11.33.$$

$$SD_2 = \sqrt{\frac{SS_2}{n-1}} = \sqrt{\frac{11.33}{5}} = 1.51.$$

Once you have computed the standard deviation for each group, compute the pooled variance. This method assumes that the variances in the two conditions are similar; that is, it is assuming homogeneity of variance). If the two populations from which these samples were drawn do in fact have similar variances, it is best to pool them when computing sampling error because you are using more data to

estimate the populations' variances. When appropriate, pooling the variances will make your significance test more accurate. The formula for the pooled variance is as follows:

$$SD_p^2 = \frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(6-1)(1.33)^2 + (6-1)(1.51)^2}{(6-1)+(6-1)} = 2.02.$$

The pooled variance formula can be a bit confusing, but it is important that you understand what it represents. The pooled variance is the average of the two sample variances weighted by the sample size. In this particular example, the sample sizes are equal (i.e., six in each group) and so you could compute the pooled variance simply by taking the average of the two variances (remember the variance is SD^2), $SD_p^2 = \frac{(1.33^2 + 1.51^2)}{2} = 2.02$. This simple formula gives the same value for the pooled variance as the more complex formula above only when the sample sizes are the same. If the sample sizes are different, the more complex formula will give a different, more accurate value because it is giving more weight to the larger sample. In general, the larger the sample, the more accurately it represents the population, and so when sample sizes are different, use the more complex formula that puts greater weight on the data from the larger sample.

The method of computing sampling error that assumes homogeneity of variance uses the pooled variance to determine the estimated standard error of the mean (SEM_i) difference, as illustrated by the following equation. Be sure to use n and not df in the denominator and note that the pooled variance (SD_p^2) is already squared.

$$SEM_i = \sqrt{\frac{SD_p^2}{n_1} + \frac{SD_p^2}{n_2}} = \sqrt{\frac{2.02}{6} + \frac{2.02}{6}} = .82.$$

The estimated standard error of the mean difference, .82, is our estimate of expected sampling error. A standard error of 0.82 indicates that the difference between the two means expected due to sampling error is 0.82.

Reading Question

11. The estimated standard error of the mean difference is
    a. used to find the critical value for the t test.
    b. an estimate of how different the two sample means are expected to be due to sampling error.
    c. Both (a) and (b) are correct.

4c. Compute the Test Statistic (Independent *t* Test)

The independent t test is the ratio of the observed deviation between the two sample means divided by the estimated standard error of the mean difference:

$$t = \frac{(M_1 - M_2)}{SEM_i} = \frac{(20.17 - 18.67)}{.82} = \frac{(1.5)}{.82} = 1.83.$$

The difference between the two means (i.e., 20.17 – 18.67 = 1.50) was 1.83 times larger than the deviation expected due to sampling error (i.e., 0.82). The critical regions of t were $t \geq 2.2281$ and $t \leq -2.2281$, and so the obtained t value did not fall in the critical region. The obtained t value is not sufficiently large to reject the null hypothesis, and so you do not reject the null hypothesis.

Reading Question

12. When the obtained t value is not further from zero, then the critical value the null hypothesis should
    a. be rejected.
    b. not be rejected.

Step 5: Compute an Effect Size and Describe It

The formula for computing the effect size of an independent t test is similar to that of a single-sample t. However, when computing the d, the denominator is the pooled variance (i.e., SD_p^2), which can be converted to the pooled standard deviation (i.e., SD_p^2) by taking its square root.[1] You computed the pooled variance in Step 3b; it was 2.02. The d computation is as follows:

$$d = \frac{\text{Observed difference between the means}}{\text{Pooled standard deviation}} = \frac{M_1 - M_2}{\sqrt{SD_p^2}} = \frac{20.17 - 18.67}{\sqrt{2.02}} = 1.06.$$

The same effect size cutoffs are used for an independent t test. If d is close to .2, the effect size is small; if it is close to .5, the effect size is medium; and if it is close to .8, it is large. An effect size of 1.06 is considered a large effect.

The results from Steps 4 and 5 may be a bit confusing. In Step 4, you failed to reject the null hypothesis, which meant that the verbal label group did not remember more of the pictures than the no verbal label group. However, the effect size of 1.06 indicates that the difference between the two means is large. Whenever the null is not rejected and yet there was a medium or large effect size, the sample size used in the study was too small for the statistical test or the effect size to be trusted. When interpreting any study's results, it is also important to consider the results from other similar studies. When Bower et al. (1975) conducted a similar study, they found that the verbal labels did improve memory performance. The combination of small sample sizes and the large effect size of your study and the fact that a similar study rejected the null suggests that you should interpret your null result with caution. In situations like this, you should obtain a larger sample size and then rerun the study.

Reading Question

13. Whenever you failed to reject the null hypothesis and yet the effect size is medium or large, you should conclude
    a. that the treatment really does work.
    b. that the treatment really does not work.
    c. that your sample size was too small and you should rerun the study with a larger sample size.

1 This is probably the most commonly used formula for computing d for independent measures designs. However, some researchers choose different denominators (Cumming, 2012). Because there are different ways to calculate d, it is important to tell the reader how you computed the effect size. In this book, we are always using the same calculation and so we do not need to say repeatedly what denominator we used, but when you present data, you should state how d was calculated.

Step 6: Interpreting the Results of the Hypothesis Test

The following paragraph summarizes these test results:

There was not a significant difference between the memory scores of those who got the verbal descriptions ($M = 20.17$, $SD = 1.33$) and those who did not ($M = 18.67$, $SD = 1.51$), $t(10) = 1.83$, $p > .05$, $d = 1.06$. However, it is important to note that the sample sizes were small, so this study should be repeated with larger sample sizes before conclusions are drawn.

Reading Question

14. When writing your results, if you failed to reject the null hypothesis and your sample size was small, you should
   a. point this out to the readers so that your report does not mislead them.
   b. not include this information in the report.

ONE-TAILED INDEPENDENT *t* TEST EXAMPLE

You and your friend Bill now turn your research efforts to studying "learning styles." Do "visual learners" have better memory for *visual* information than "verbal learners"? If the learning styles theory is accurate, "visual learners" should recall more visual information than "verbal learners." Given that the theory makes a specific prediction, a one-tailed *t* test is appropriate in this situation. Twenty-nine "verbal learners" and 31 "visual learners" volunteered to participate in a study investigating this question. All learners were presented with simple line drawings and then were asked to re-create as many of the line drawings as they could remember. The visual memory scores of each group are listed below. You correctly use a one-tailed independent *t* test with an alpha level of .05 to determine if "visual learners" recall more of the line drawings than "verbal learners."

Group 1: Verbal Learners Group:

$$M_1 = 15.00;\ SD_1 = 1.41,\ n_1 = 29$$

Group 2: Visual Learner Group:

$$M_2 = 15.25;\ SD_2 = 1.67,\ n_2 = 31$$

Step 1: Examine the Statistical Assumptions

As in the previous example, these data meet all of the statistical assumptions. The data within each condition are independent, and the DV is measured on an interval/ratio scale. You also know that memory scores tend to be normally distributed, so the normality assumption is likely met. As before, the homogeneity of variance assumption will be assessed with Levene's test provided by SPSS.

Step 2: State the Null and Research Hypotheses Symbolically and Verbally

The research hypothesis for a one-tailed test specifies which of the two sample means will be higher if the "treatment" works. For this example, Group 1 was "verbal learners," and Group 2 was "visual learners." The learning styles theory predicts that "visual learners" should learn visual information better than "verbal learners" do, and this prediction is represented symbolically by the following

Table 10.2 Symbolic and Verbal Representations of One-Tailed Research and Null Hypotheses for an Independent *t* Test

	Symbolic	*Verbal*	*Mean Difference Created by*
Research hypothesis (H_1)	H_1: $\mu_1 < \mu_2$ OR H_1: $\mu_1 - \mu_2 < 0$	The population of "visual learners" *does* learn visual information better than the population of "verbal learners" does.	Visual learning style being better at learning visual information
Null hypothesis (H_0)	H_0: $\mu_1 \geq \mu_2$ OR H_0: $\mu_1 - \mu_2 \geq 0$	The population of "visual learners" *does not* learn visual information better than the population of "verbal learners" does.	Sampling error

research hypothesis: $\mu_{verbal} < \mu_{visual}$ or $\mu_{verbal} - \mu_{visual} < 0$. The null hypothesis is the opposite of the research hypothesis, including all other possible outcomes. Thus, the null hypothesis states that $\mu_{verbal} \geq \mu_{visual}$ or $\mu_{verbal} - \mu_{visual} \geq 0$. In other words, the null hypothesis is that "visual learners" do not learn visual information any better than the "verbal learners." The research and null hypotheses for this one-tailed independent *t* test are shown in Table 10.2.

Reading Question

15. Which of the following *could* represent a null hypothesis when doing a one-tailed independent test? Note that this question is not asking about the study described in the text. Which of the following could possibly be a one-tailed null hypothesis?

a. $\mu_1 \geq \mu_2$
b. $\mu_1 < \mu_2$
c. $\mu_1 > \mu_2$
d. $\mu_1 = \mu_2$

Step 3: Compute the Degrees of Freedom and Define the Critical Region

Computing the *df* for one-tailed tests is done in exactly the same manner as with a two-tailed test. In this case,

$$df = (n_1 - 1) + (n_2 - 1) = (29 - 1) + (31 - 1) = 58.$$

The one-tailed *t* table indicates that a study with a $df = 58$, when using $\alpha = .05$, has a critical value of 1.6716. The research hypothesis predicts that μ_1, the verbal group, will be less than μ_2, the visual group, so it is predicting a negative obtained *t* value, $\mu_{verbal} - \mu_{visual} < 0$. Thus, the critical region is in the negative side of the distribution. The null should be rejected if $t \leq -1.6716$. If "verbal learners" were labeled as Group 2 and "visual learners" as Group 1, we would have expected a positive *t*, and the critical region would have been in the positive side of the distribution.

Reading Question

16. When you are doing a one-tailed *t* test, the critical region is always on the side of the *t* distribution that is predicted by the
   a. null hypothesis.
   b. research hypothesis.

Step 4: Compute the Test Statistic

4a. Compute the Deviation Between the Two Sample Means

Again, the numerator of the test is the difference between the two sample means:

$$(M_1 - M_2) = (15.00 - 15.25) = -0.25.$$

4b. Compute the Average Sample Error That Is Expected

As in the previous example, you need the standard deviation for *each group of scores*. In this problem, they were provided for you, but you should know how to compute them if you need to for future problems. Review the previous example if you are unsure how this is done. Once the standard deviation for Group 1 (i.e., $SD_1 = 1.41$) and the standard deviation for Group 2 (i.e., $SD_2 = 1.67$) have been computed, you compute the pooled variance (i.e., SD_p^2) as follows:

$$SD_p^2 = \frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(29 - 1)(1.41)^2 + (31 - 1)(1.67)^2}{(29 - 1) + (31 - 1)} = 2.402.$$

The pooled variance is the average variance for the two samples weighted by the sample size. In this case, the second sample was a bit larger than the second sample, and so the variance associated with that sample was given more weight when computing the pooled variance.

It only makes sense to pool (or average) the two standard deviations if they are estimating the same population parameter. If the standard deviations are significantly different from each other, the pooled variance must be computed in a different way. Fortunately, SPSS does all of these computations automatically, and so we will revisit this issue in the SPSS section that follows.

After you have computed the pooled variance, you will use it to compute the estimated standard error of the mean difference, which is done as follows:

$$SEM_i = \sqrt{\frac{SD_p^2}{n_1} + \frac{SD_p^2}{n_2}} = \sqrt{\frac{2.402}{29} + \frac{2.402}{31}} = 0.40.$$

Be sure to use *n* from each group and not *df* in the denominator when computing the estimated standard error of the mean difference. Also, note that the pooled variance (SD_p^2) is already squared, and so you should not square it again.

Reading Question

17. When computing the standard error of the mean difference, the equation calls for using
   a. the degrees of freedom for each group.
   b. the sample sizes from each group.

4c. Compute the Test Statistic (Independent *t* Test)

Again, the *t* test computation is identical to a two-tailed test:

$$t = \frac{(M_1 - M_2)}{SEM_i} = \frac{(15.00 - 15.25)}{0.40} = \frac{-0.250}{0.40} = -0.63.$$

The obtained *t* value of −0.31 is not further from zero than the critical value of −1.6716. Therefore, you do not reject the null hypothesis.

Reading Question

18. Which of the following values can never be negative?
    a. The numerator of a *t* test
    b. The obtained *t* value
    c. The standard error of the mean difference

Step 5: Compute an Effect Size and Describe It

Again, computing the effect size for one- and two-tailed tests is identical:

$$d = \frac{\text{Observed difference between the means}}{\text{Standard deviation}} = \frac{M_1 - M_2}{\sqrt{SD_p^2}} = \frac{15.00 - 15.25}{\sqrt{2.402}} = -0.16.$$

An effect size with an absolute value of 0.16 is small. It means that the learning style of the learners had virtually no effect at all on memory scores.

Reading Question

19. When determining if an effect size is small, medium, or large, you should
    a. ignore the sign of the computed effect size and use its absolute value.
    b. recognize that negative effect sizes are always small effect sizes.

Step 6: Interpreting the Results of the Hypothesis Test

When interpreting the results of any study, you should always consider the results of the significance test *and* the effect size. In this learning styles study, the significance test suggested that the difference in mean memory scores between visual and verbal learners was probably just sampling error. In other words, the data did not support the learning styles theory. The small effect size in this study also did not support the idea that learning styles influence memory performance. This study generated an unambiguous "null result." If many other studies in the literature report similar null results, then you can be more confident that your null result accurately reflects the true situation.

Even though you may have heard a lot about learning styles and have been told that you are a "visual" or "auditory" or "verbal" learner, you should know that experimental studies have generally found null results when investigating learning styles (Cook, Gelula, Dupras, & Schwartz, 2007; Pashler, McDaniel, Rohrer, & Bjork, 2008). In fact, the collection of null results is sufficiently large that memory researchers generally agree that, currently, there is no evidence supporting learning styles theory (Pasher et al., 2008). Instead, memory researchers suggest that instructors should incorporate

memory strategies supported by evidence. A great deal of memory research supports the effectiveness of using deeper levels of processing or retrieval practice effects. In other words, you should think about the meaning of what you are trying to learn and practice recalling the information from your memory. Research suggests that these studying strategies will increase your academic performance and that relying on learning styles will not.

The results from your and Bill's learning styles study might be summarized as follows:

Contrary to the prediction of learning styles theory, the visual learners ($M = 15.25$, $SD = 1.41$) did not learn significantly more visual information than the verbal learners ($M = 15.00$, $SD = 1.67$), $t(58) = -0.63$, $p > .05$ (one-tailed), $d = -0.16$. The study's null result is consistent with the null results found by several other researchers investigating learning styles. Collectively, the null results of several studies on learning styles strongly suggest that there is very little, if any, merit in the learning styles theory of learning.

Reading Question

20. Researchers need to be very cautious when interpreting null results because
    a. they prove that the IV had no effect on the DV.
    b. a null result might occur because of a problem with the study's experimental procedure.

OTHER ALPHA LEVELS

In both of the previous examples, you used an alpha of .05. If you were to use an alpha of .01, rather than .05, it would be harder to reject the null hypothesis. This means that you would have less statistical power but a lower risk of making a Type I error.

Reading Question

21. Which alpha value has a lower risk of making a Type I error?
    a. .05
    b. .01

Reading Question

22. Which alpha value has a lower risk of making a Type II error?
    a. .05
    b. .01

SPSS

Data File

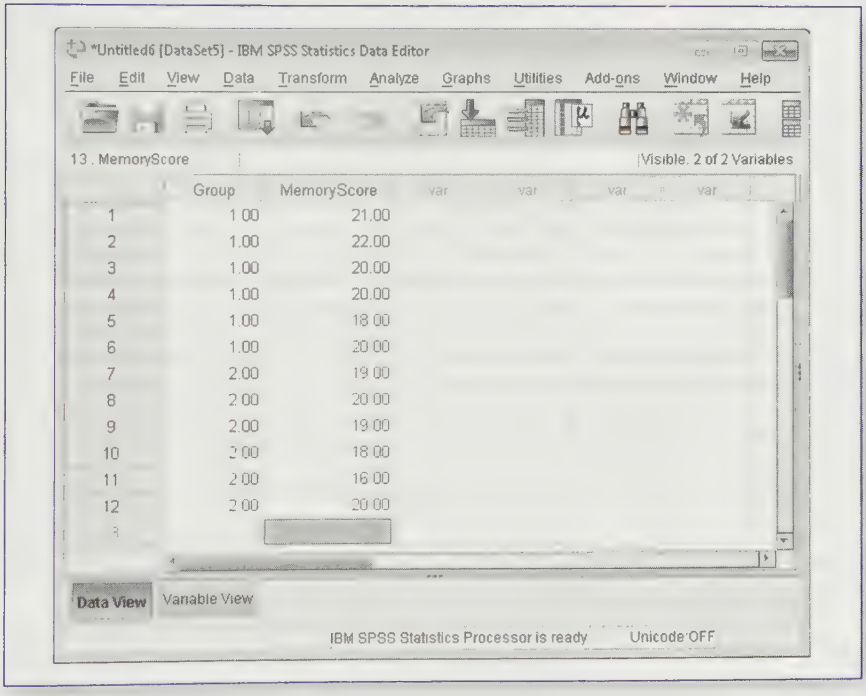
We are going to use the data from the first example in the chapter to illustrate how to use SPSS to conduct an independent samples t test. The data are reproduced as follows:

Group 1: With Verbal Descriptions Group: 21, 22, 20, 20, 18, 20

Group 2: Without Verbal Descriptions Group: 19, 20, 19, 18, 16, 20

To enter data for an independent samples *t* test, you will need two columns. The first is the IV (i.e., grouping variable). In this case, we need to enter a number to indicate the group that each person was in (i.e., verbal descriptions or no verbal descriptions). You can use any numbers, but below we used a "1" to indicate verbal descriptions and a "2" to indicate no verbal descriptions. The second column is the DV (the dependent variable, i.e., the variable on which the two groups are being compared). In this case, the DV is the visual memory score of each person. When you are done, the data file should look like Figure 10.4.

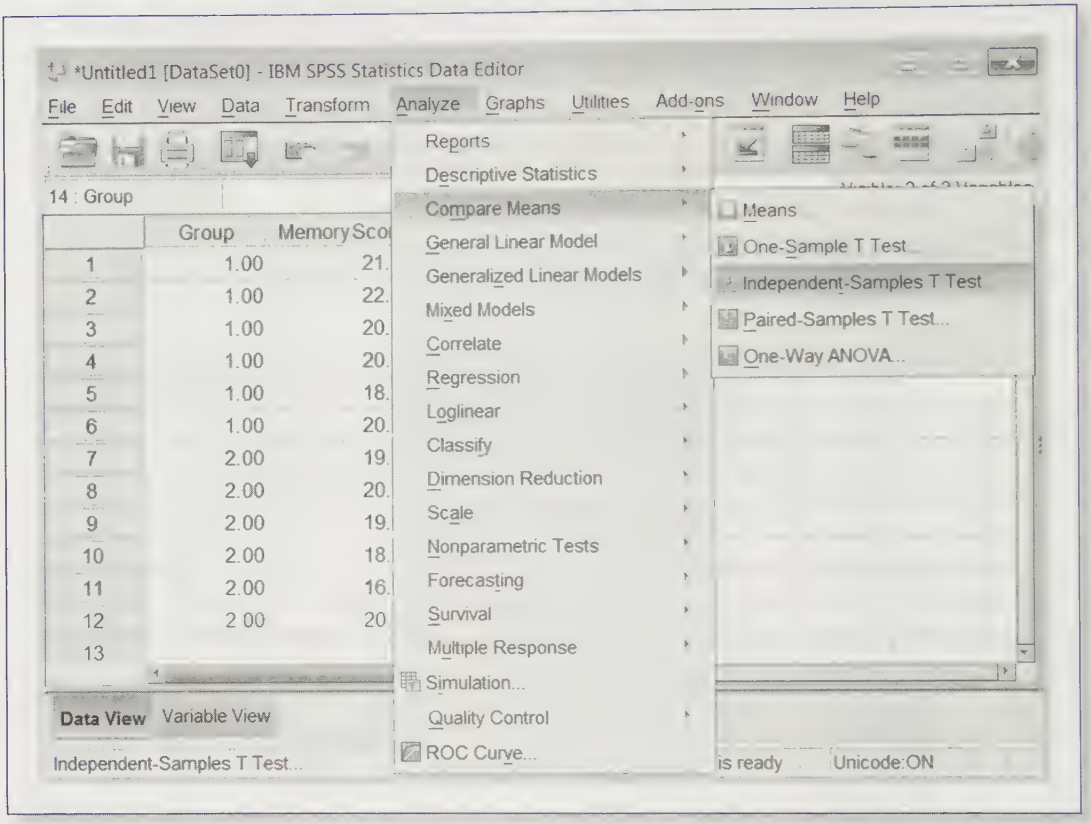
Figure 10.4 SPSS Screenshot of Data Entry Screen for an Independent *t* Test



Computing an Independent Samples *t* Test

- Click on the Analyze menu. Choose Compare Means and then Independent Samples *t* Test (see Figure 10.5).
- Move the Independent Variable (the one that indicates which group someone is in) into the Grouping Variable box, and click on Define. Enter the values you used to designate Group 1 and Group 2 in the appropriate boxes (in the above screenshot, you would enter the values 1 and 2, respectively).

Figure 10.5 SPSS Screenshot of Choosing an Independent *t* Test



- Move the Dependent Variable (the one that indicates the actual scores of the participants) into the Test Variables box (see Figure 10.6).
- Click on the OK button.

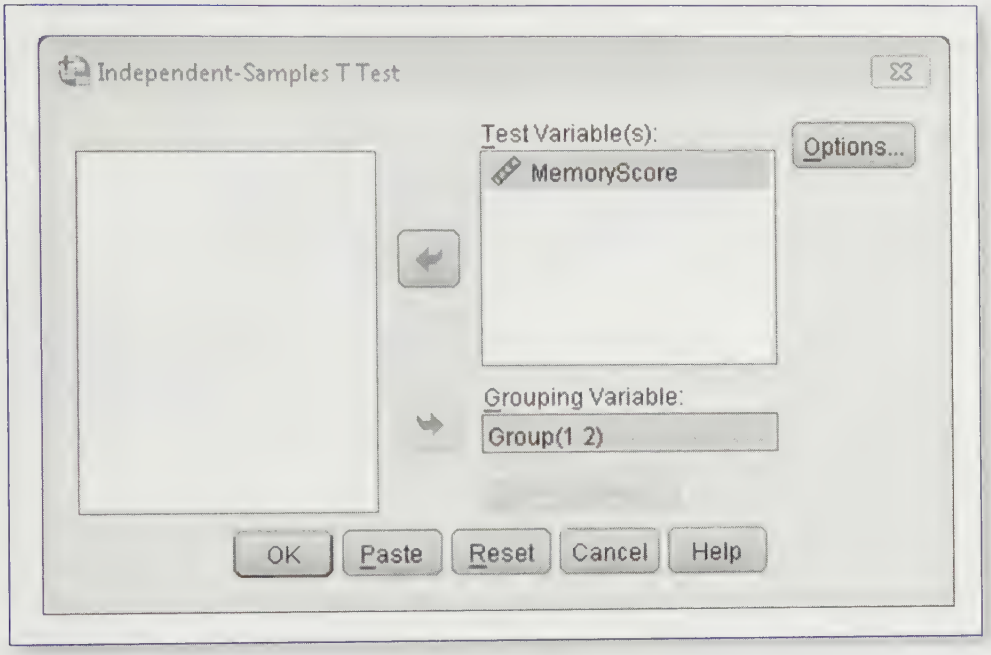
Output

The SPSS output for this analyses is given in Figure 10.7.

Levene's Test (Test for Homogeneity of Variance)

There are two ways to estimate expected sampling error (i.e., the denominator of the independent *t* test). One way assumes homogeneity of variance, and the other does not. This assumption states that the variances in the two populations (e.g., verbal descriptions and no verbal descriptions) are the same. This assumption is important because if the variances are equal, you should compute the amount of expected sampling error as illustrated in the previous examples. However, if this assumption is

Figure 10.6 SPSS Screenshot of the IV and the DV for an Independent *t* Test



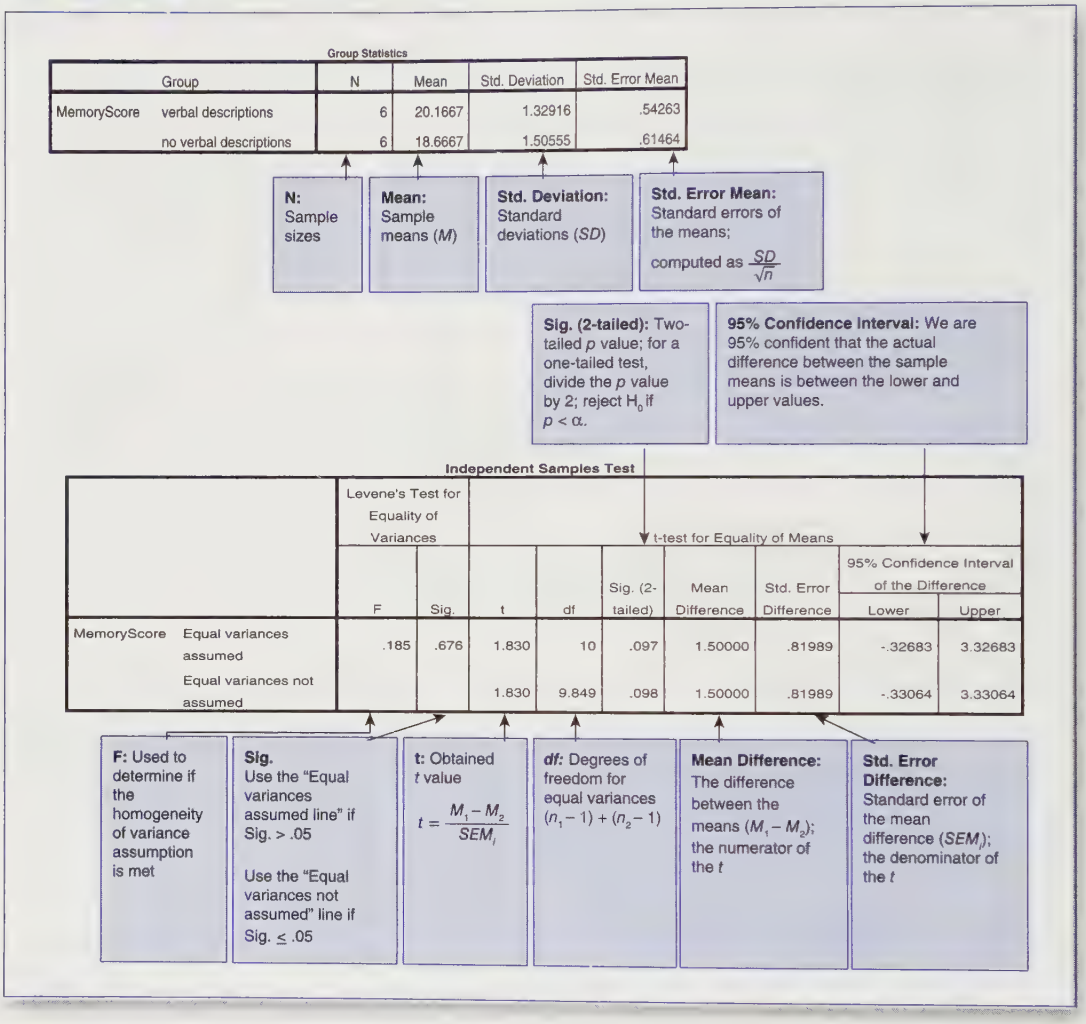
violated, the estimated sampling error should be computed differently. Levene's test will indicate which method for computing sampling error is most appropriate. Just as we do a *t* test to determine if two means are significantly different, we can do a test to determine if the two variances are significantly different. The standard deviations for the verbal learners and visual learners were 1.41 and 1.67, respectively. Thus, the variance was 1.41^2 or 1.99 for the verbal learners and 1.67^2 or 2.79 for the visual learners. SPSS automatically runs Levene's test for homogeneity of variance to determine if 1.99 is significantly different from 2.79. Note that our double standard deviation rule suggests that the homogeneity of variance assumption is not violated.

Reading Question

23. Levene's test is automatically run by SPSS to determine if the _____ of the two groups are significantly different.
   a. means
   b. variances

If the variances of the two groups are not significantly different, the proper way to compute the estimate of sampling error is to use the "Equal variances assumed" method. If the variances are significantly different, the proper way to compute the estimate of sampling error is to use the "Equal variances not assumed" method. You can determine if the variances are equal or not by looking at the "Sig." value under the "Levene's Test for Equality of Variances" label in the "Independent Samples Test" output. If the Sig. value is less than or equal to .05, the variances are not similar and the equal variance

Figure 10.7 Annotated SPSS Output for an Independent *t* Test



assumption was violated. If the Sig. value is greater than .05, the variances are similar and the assumption of equal variance was met.

Reading Question

24. Use the "Independent Samples Test" output to find the Sig. value for Levene's test and then determine if the variances are similar or not similar. The Sig. (*p*) value for Levene's test is _____ and therefore the variances for the verbal and visual groups _____.

a. .676; are similar (i.e., equal)

b. .097; are not similar (i.e., not equal)

Levene's test indicated that the variances in the two groups were equal. Therefore, the best way to compute the estimate of sampling error is to use the "Equal variances assumed" method. If you look in the left-most cell of the "Independent Samples Test" output, you will see two labels. The one at the top is "Equal variances assumed," and the one at the bottom is "Equal variances not assumed." Levene's test indicated that we should use the "Equal variances assumed" method. SPSS automatically computes two different *t* tests—one using the "equal variances" method and the other using the "not equal variances" method. The results of both *t* tests are shown in the output. In this case, we should choose the obtained *t* value and degrees of freedom[2] that is across from the "Equal variances assumed" heading (i.e., $t = -1.830$ and $df = 10$).

Reading Question

25. If the Sig. for Levene's test was less than .05, the correct *df* would have been
    a. 10
    b. 9.849

Reading Question

26. When entering the data for an independent *t* test, you should have
    a. a column for each person.
    b. a column for the IV and a column for the DV.

Reading Question

27. Use the "Group Statistics" output to determine the mean and standard deviation of the verbal descriptions group. The mean and standard deviation of the verbal descriptions group were
    a. 20.1667 and 1.32916, respectively.
    b. 18.6667 and 1.50555, respectively.

Reading Question

28. The *p* value for this *t* test was
    a. .676.
    b. .097.

Reading Question

29. To get a one-tailed *p* value, you must
    a. divide the two-tailed *p* value in half.
    b. multiply the two-tailed *p* value by 2.

2 Like the obtained *t* value, the degrees of freedom are also computed differently when the homogeneity of variance assumption is violated. You will not need to do this by hand in this book, but the formula is different from the one you use when the homogeneity of variance assumption is met. The *df* formula is $df = \frac{\left(\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}\right)^2}{\frac{\left(SD_1^2/n_1\right)^2}{n_1+1} + \frac{\left(SD_2^2/n_2\right)^2}{n_2+1}} - 2.$ When the variances are not similar, the SEM_i is computed as $\sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}$. Finally, when the *t* test is modified in this way, it is called a Welch *t* test (Welch, 1947). Fortunately, SPSS will do all of these calculations automatically. All you have to know is which *t* test to interpret.

OVERVIEW OF THE ACTIVITIES

In Activity 10.1, you will work through all of the steps of hypothesis testing for a one-tailed hypothesis test. While doing so, you will work to understand conceptually what you are doing at each step and the implications of the statistical decisions you make with regards to Type I error, Type II error, and statistical power. In this activity, you will also review information from two studies and compare and contrast the results of the significance test, effect size, and study design. In Activity 10.2, you will review some of the material from Activity 10.1, do a two-tailed hypothesis test, and work with SPSS. In Activity 10.3, you will read different research scenarios and decide which statistic should be used to test the hypothesis. Activity 10.4 is a group exercise giving you the opportunity to collect data using an independent measures design, a repeated measures design, and a matched design. This activity is intended to help you understand the relative advantages/disadvantages of the different types of designs and to review the information from Chapters 9 and 10. Finally, in Activity 10.5, you will compute and interpret confidence intervals for a two-group independent measures design.

Activity 10.1: Hypothesis Testing With the Independent *t* Test

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- State null and research hypotheses for an independent *t* test
- Compute and interpret the results of an independent *t* test
- Explain how sampling error influences an independent *t* test
- Compute and interpret the effect size estimate for an independent *t* test
- Use the distributions of sample means to locate the probabilities of Type I error, Type II error, statistical power, and the probability of rejecting a false null hypothesis

INDEPENDENT *t* TEST EXAMPLE

It is well known that acetaminophen lessens people's physical pain. A recent study suggests that acetaminophen can lower people's psychological pain as well (DeWall et al., 2010). Surprised by these findings, you decided to see if acetaminophen works for different types of psychological pain than those assessed in the original research. You obtained a sample of volunteers and gave half of them acetaminophen and the other half a placebo pill. Both groups of participants read socially painful stories and rated how painful the experience would be for them. Specifically, you want to know if the mean social pain rating of the acetaminophen group is statistically lower than the mean social pain rating of the control group. Stated differently, you want to know if the mean difference in social pain rating between these two samples is likely or unlikely to be due to sampling error. You are

predicting that the acetaminophen will *reduce* social pain ratings; therefore, you used a one-tailed hypothesis test with $\alpha = .05$.

$n_{drug} = 31$, $M_{drug} = 213$, $SD_{drug} = 18$.

$n_{control} = 31$, $M_{control} = 222$, $SD_{control} = 20$.

STATISTICAL ASSUMPTIONS

1. Match the assumption to the fact that is relevant to that assumption.

   _____ Independence

   _____ Appropriate measurement of the IV and the DV

   _____ Normality

   _____ Homogeneity of variance

   a. Samples of this size tend to form distributions of sample means with a normal shape.
   b. Data were collected from one participant at a time.
   c. This assumption will be assessed later by Levene's test.
   d. The IV manipulation is well defined, and the participants' responses were given on an interval/ratio scale.

2. How do you know that the homogeneity of variance assumption is not violated?
   a. One standard deviation is not double the size of the other.
   b. The standard deviations are different by less than 5.

3. How do you know that the normality assumption is probably not violated?
   a. The sample size in each group is greater than 30.
   b. The research scenario indicates that social pain scores are normally distributed.

UNDERSTANDING THE NULL AND RESEARCH HYPOTHESES

4. Write H_0 next to the symbolic notations for the null hypothesis and H_1 next to the research hypothesis.

   a. _____ $\mu_{drug} = \mu_{placebo}$
   b. _____ $\mu_{drug} \neq \mu_{placebo}$
   c. _____ $\mu_{drug} > \mu_{placebo}$
   d. _____ $\mu_{drug} > \mu_{placebo}$
   e. _____ $\mu_{drug} < \mu_{placebo}$
   f. _____ $\mu_{drug} < \mu_{placebo}$

5. Write H_0 next to the verbal description of the null hypothesis and H_1 next to the research hypothesis.
   a. _____If a population of people were given the drug, they would have higher social pain scores than would a population of people given the placebo.
   b. _____If a population of people were given the drug, they would not have higher social pain scores than would a population of people given the placebo.
   c. _____If a population of people were given the drug, they would have lower social pain scores than would a population of people given the placebo.
   d. _____If a population of people were given the drug, they would not have lower social pain scores than would a population of people given the placebo.
   e. _____If a population of people were given the drug, their social pain scores would be different from a population of people given the placebo.
   f. _____If a population of people were given the drug, their social pain scores would not be different from a population of people given the placebo.

6. Assuming that the null hypothesis is true (i.e., that acetaminophen does not affect social pain at all), what precise value would you expect for the *mean difference* between the mean ratings of social pain for the acetaminophen sample and the placebo sample?

   Precise expected value for the *mean difference* between the two sample means if the null is true = _____.

7. In this situation, you obtained two samples of people from the population. One sample was assigned to take acetaminophen, and the other sample was assigned to take a placebo. Even if you assume that the acetaminophen does not affect social pain at all, would you be surprised if the samples' *mean difference* was not *exactly* the value you indicated earlier?
   a. No, I would not be surprised. The two sample means are likely to be different because the population parameters are different.
   b. No, I would not be surprised. The two sample means are likely to be different because of sampling error even if the drug has no effect.
   c. Yes, I would be surprised. The two sample means should be exactly the same if the drug does not affect social pain.
   d. Yes, I would be surprised. The two sample means should be exactly the same because the sizes of the two samples are the same.

8. Just as was the case with the z for a sample mean, the single-sample t test, and the related samples t test, you must compute the amount of sampling error that is "expected by chance or sampling error" before you can determine if the null hypothesis is likely or unlikely to be true. The acetaminophen and control groups each had 31 people with social pain SDs of 18 and 20, respectively. Use the two equations provided below to compute the value for standard error of the mean difference, estimated sampling error.

$$SD_p^2 = \frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{(n_1 - 1) + (n_2 - 1)} =$$

$$SEM_i = \sqrt{\frac{SD_p^2}{n_1} + \frac{SD_p^2}{n_2}} =$$

9. What does the SEM_i estimate?
   a. The typical amount of sampling error in the two separate samples
   b. The typical distance between the two sample means expected due to sampling error
   c. The typical distance between the sample scores and the means for the two samples

10. Assuming that acetaminophen does not affect social pain at all (i.e., assuming the null hypothesis is true), create a distribution of sample *mean differences* based on the two sample sizes of $n_{drug} = 31$ and $n_{control} = 31$. This distribution represents the frequency of all possible sample *mean differences* when both sample sizes are 31. You should label the frequency histogram so that it is centered on the sample *mean difference* you expect if the null is true. Look back at your answer to Question 6 to determine what the mean of the distribution of sample means should be if the null hypothesis is true.

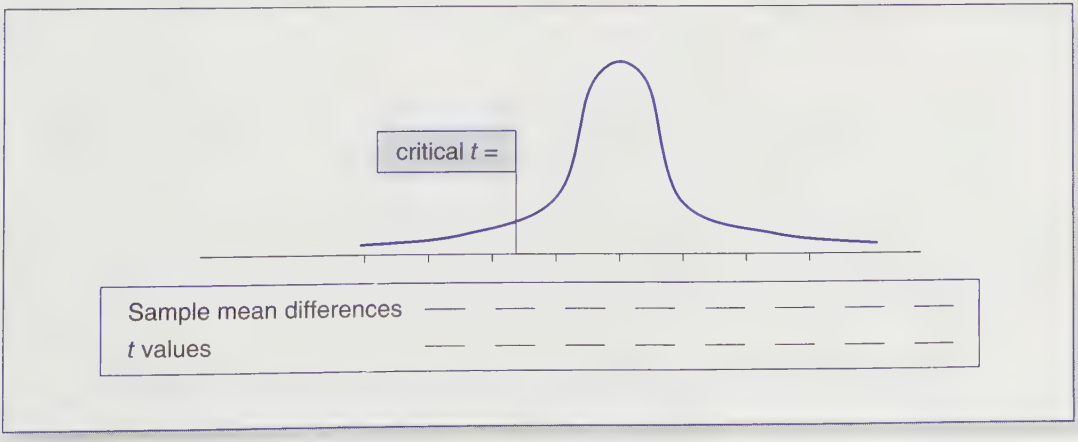
    After you label the mean, label each standard error of the mean to the right and left of the mean. You computed the standard error of the mean in Question 8.

    Next, use the independent *t* formula to determine the *t* score associated with each sample mean difference. In other words, convert each raw mean difference score on the distribution into a *t* value.

    Finally, label the critical *t* value and shade the critical region. You will need to consult a critical *t* value table to locate the critical value of *t*.

Evaluating the Likelihood of the Null Hypothesis

11. Now that the critical region of *t* is determined, you can find the *t* value associated with the *mean difference* between the samples and determine if it is in the critical region of the *t* distribution (also known as the region of rejection). The sample means were $M_{drug} = 213$ and $M_{control} = 222$. Locate the mean difference on the distribution of sample mean differences above. Is it in the critical region or outside of the critical region?
    a. In the critical region
    b. Outside of the critical region



12. Compute the obtained t value.

13. What should you conclude about the null hypothesis and why?
    a. Reject H_0 because the computed t value is in the critical region
    b. Reject H_0 because the computed t value is outside of the critical region
    c. Fail to reject H_0 because the computed t value is in the critical region
    d. Fail to reject H_0 because the computed t value is outside of the critical region

Effect Size

The process of significance testing you just completed in the previous question tells us whether or not the null hypothesis is likely to be true. It does not indicate how effective the IV was in affecting the DV. In the preceding scenario, the null hypothesis was rejected, but the researchers only know that acetaminophen will probably decrease social pain. They do not know how effective it would actually be. To get this information, they must compute an estimate of the study's effect size.

14. Compute the estimate of effect size (d) of acetaminophen on social pain. (Note: When computing the d for an independent t test, you need to use the "pooled standard deviation" or the square root of the pooled variance. Consult your reading for a reminder.)

15. Is the effect size small, small to medium, medium, medium to large, or large?

Summarize the Results

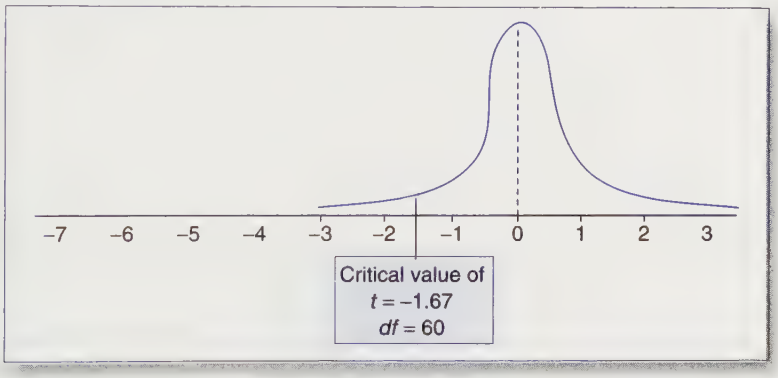
16. Choose the best APA style summary of the results.
    a. The mean social pain ratings of the acetaminophen group ($M = 213$, $SD = 18$) and the placebo group ($M = 222$, $SD = 20$) were significantly different, $t(60) = -1.86$, $p < .05$, $d = .47$. The effect size suggests that acetaminophen is very effective at reducing social pain.
    b. The mean social pain ratings of the acetaminophen group ($M = 213$, $SD = 18$) were not significantly different from the mean ratings of the placebo group ($M = 222$, $SD = 20$), $t(60) = -1.86$, $p > .05$, $d = .47$.
    c. The mean social pain ratings of the acetaminophen group ($M = 213$, $SD = 18$) were significantly less than the mean ratings of the placebo group ($M = 222$, $SD = 20$), $t(60) = -1.86$, $p < .05$, $d = .47$. The effect size suggests that acetaminophen is moderately effective at reducing social pain because it reduces it by about a half a standard deviation.

Type I Errors, Type II Errors, and Statistical Power

17. *Before* collecting data from the sample that took acetaminophen, you can't locate the center of the distribution of sample mean differences if the research hypothesis is true. However, *after* you have the data from the sample, you do have *an idea* of where the research hypothesis distribution of sample mean differences is located. What can you use to estimate the location of this distribution? Provide a specific value. (Hint: What value might represent the *mean difference* between the social pain scores of all those who took acetaminophen and all those who took the placebo if the research hypothesis is true?)

    Specific value for the center of the research hypothesis distribution of sample *mean differences* = _____.

    A major benefit of locating the center of the research hypothesis distribution of sample mean differences, even if it is an *estimated* center that is inferred from sample statistics, is that it allows researchers to quantify several other very important statistical concepts. This quantification process can be illustrated by "building a distribution." You have already built one of the distributions that are necessary in Question 10. As you know from Question 10, the null hypothesis distribution of sample mean differences is centered at a *t* value of 0. The null *t* distribution is re-created for you as follows:



18. The second distribution you need to build is the distribution of the sample mean differences if the research hypothesis is true. As mentioned earlier, the center of this distribution is determined by the actual mean difference of the samples. In this case, the actual sample mean difference was −9, meaning that the center of this distribution of sample means is at −9 on the raw score number line. What is the *t* value associated with a sample mean difference of −9?

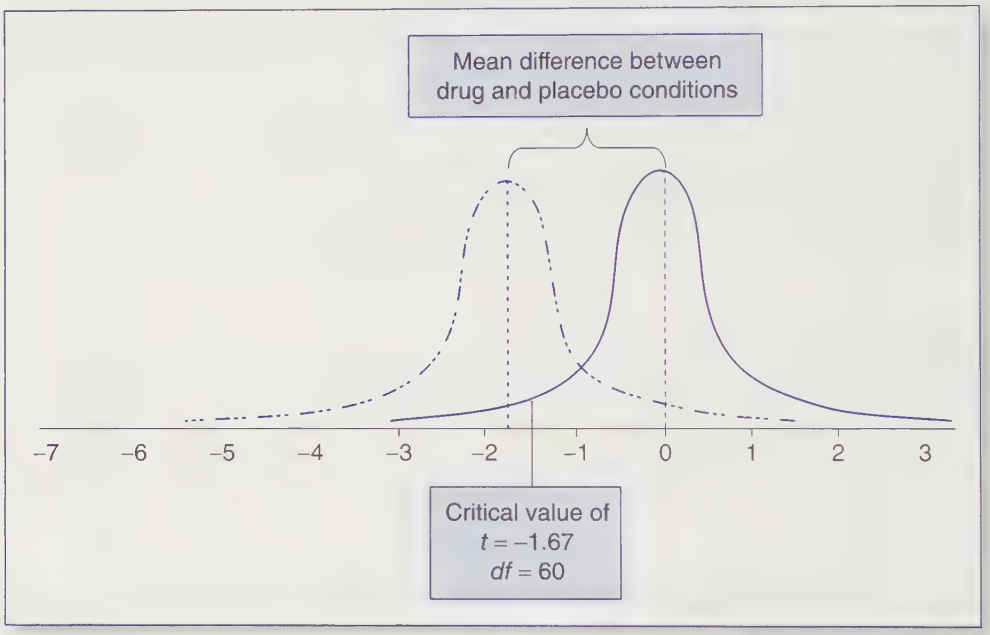
    Draw a vertical line on the above figure at the *t* value associated with a mean difference of −9 to represent the center of the distribution of sample means if the research hypothesis is true. Remember that the only reason we can locate this curve's center is because you know the actual mean difference of the samples. You can't know this *before* you collect data.

19. In Activity 6.1, you created the research hypothesis distribution of sample means by sketching a normally shaped distribution. When working with *t* distributions, the research distribution of sample means can be skewed with smaller sample sizes.[3] Here, our sample size is large enough

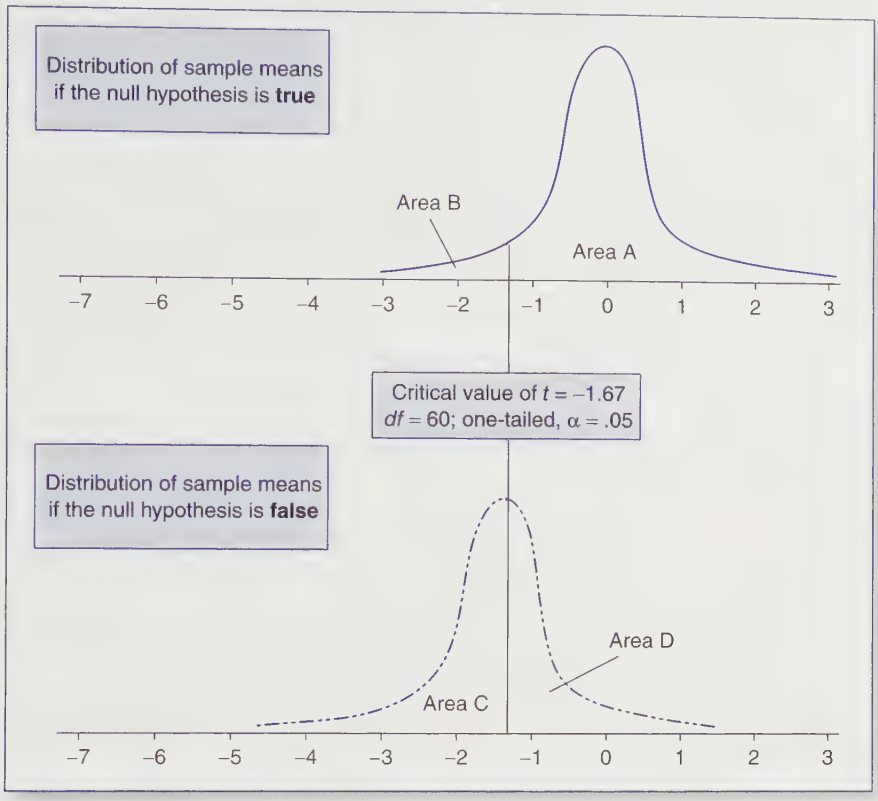
3 The severity of the research curve's skew depends on the effect size and the sample size, but *as sample size increases, the research curve approaches a normal shape* (Cummings, 2012).

that the distribution is essentially normal. In addition, we are less concerned with the research curve's precise shape than we are with using it to locate the Type II and statistical power regions in the research curve. Therefore, sketch in the distribution of sample mean differences if the research hypothesis is true, assuming it has a normal shape. (Just try to duplicate the null curve's shape and spread but have it centered at −1.86 on the *t* score number line rather than at 0, where the null curve is centered.)

After you have completed sketching in the research hypothesis distribution of sample mean differences, look at the following figure and confirm that your figure looks like it. If it doesn't, determine what is wrong and fix it. Be sure you understand why the figure looks as it does.



Just as can be done with the *z* for the sample mean, single-sample *t* test, and the related samples *t* test, we can use these curves to quantify the probability of important statistical concepts. The null distribution of sample mean differences represents all possible outcomes if the null hypothesis is true. The research distribution of sample mean differences represents all possible outcomes if the research hypothesis is true. By "cutting" these curves into different sections at the critical value of *t*, we can estimate the probability of (a) rejecting a false null (i.e., statistical power), (b) failing to reject a false null (i.e., Type II error), (c) rejecting a true null (i.e., Type I error), and (d) failing to reject a true null. In the following figure, there is a distribution of sample mean differences for the null hypothesis and a distribution of sample mean differences for the research hypothesis. The two curves have been "separated" so that it is easier to see the distinct sections of each curve. The areas under each of these respective curves represent statistical power, Type II errors, Type I errors, and failing to reject a true null. Try to determine which areas under each curve represent what in the following figure.



20. Match the area to the statistical outcome

Type I error [rejecting a true null]: Area _____.

Type II error [failing to a false null]: Area _____.

Statistical power [rejecting a false null]: Area _____.

Failing to reject a true null: Area _____.

21. Go back to Activity 6.1. Make note of the fact that the areas under the curve that correspond to each of the above concepts are not the same as in Activity 6.1. What is different about this scenario that changed the locations of the above areas on the curves?

a. In this scenario, the hypothesis test was one-tailed, while in the previous scenario, the hypothesis test was two-tailed.

b. In this scenario, the research hypothesis predicted a negative difference, while in the previous scenario, the predicted mean difference was positive.

c. In this scenario, there were two samples that resulted in two separate curves, while in the previous scenario, there was just one sample and one curve.

22. If a one-tailed, $\alpha = .01$ significance test was used instead of a one-tailed, $\alpha = .05$ significance test, what would happen to the critical value line in the preceding graphs? It would move
    a. closer to zero.
    b. closer to the left tail.

23. What would happen to the probability of each of the following if a one-tailed, $\alpha = .01$ significance test was used instead of a one-tailed, $\alpha = .05$ significance test? Indicate if each would increase, decrease, or stay the same.

Type I error:	Increase	Decrease	Stay the same
Type II error:	Increase	Decrease	Stay the same
Statistical power:	Increase	Decrease	Stay the same

24. Which two of the following values are estimates and which one value is known precisely?

Type I error:	Estimate only	Precise value is known
Type II error:	Estimate only	Precise value is known
Statistical power:	Estimate only	Precise value is known

25. Why are the two values you identified in Question 24 estimates?
    a. Type II error and power are estimated based on the null hypothesis curve.
    b. The mean (center) and the standard deviation (spread) of the research distribution curve are estimated from the sample data.

Summary of Key Points in Significance Testing

26. So far in this course, we have gone through several examples illustrating the details of how significance testing works. These details included each of the following key points. Read through these key points and try to fill in the key terms that are missing. You may use some of the terms from page 345 more than once.

    - The _____ allows us to predict the center, shape, and spread of the distribution of sample means (or the distribution of sample *mean differences*) if the null is true.
        - State the theorem below.

    - The _____ hypothesis precisely locates the center of the null distribution of sample means (or the null distribution of sample mean differences). This location is at a z value of _____ or at t value of _____.
    - The center of the _____ hypothesis distribution of sample means (or mean differences) is estimated from the sample mean (or sample mean difference) and cannot be known before data are collected.

- The probability of a _____ is set by researchers when they set the α value.
- By building the null and research hypothesis distributions of sample means (or sample mean differences), we can quantify the probability of failing to reject a false null (i.e., _____ _____), as well as the likelihood of rejecting a false null (i.e., _____).
- The _____ of *t* or *z* "cuts" likely values if the null is true from unlikely values if the null is true; if a statistic (e.g., *z* value, *t* value) is more extreme than this value, the null hypothesis is unlikely to be true.
- If the null is rejected, the _____ hypothesis is considered likely to be true.
- After a _____ is performed, researchers know whether or not the null is likely to be true. They do not know how effective the IV was at affecting the DV. To quantify the impact of the IV on the DV, researchers must compute a/an _____.

Type I error	0	Statistical power
Type II error	Power	Central limit theorem
Null	Effect size	Significance test
Critical value	Research	

27. A researcher obtained the scores for all 1.5 million students who took the SAT this year and found that males scored significantly higher than females ($p < .001$). Given the sample size, what additional information would you want before you decide if this statistically significant difference is large enough to be an important difference?
    a. The confidence interval
    b. The effect size

Putting It All Together

Suppose that you have a friend who suffers from debilitating panic attacks that have not been helped by treatments that are currently on the market. Your friend's psychiatrist asks if he would like to participate in a clinical trial. There are two clinical trials available (Drug A or Drug B), and your friend asks you for help in deciding which drug seems more promising based on the available data.

Although neither of the drugs are yet approved by the Food and Drug Administration (FDA), both are currently in the human trial stage of the approval process. The side effects for both drugs seem to be minor, and if your friend agrees to participate in one of the trials, all costs would be paid by a National Institutes of Health (NIH) research grant. The preliminary results from the trials are described below.

Drug A information: In total, 250 participants took Drug A for 3 months, while another 250 participants took a placebo for 3 months in a double-blind study. While taking the "drug," the participants recorded the number of panic attacks they experienced. The mean number of panic attacks in the Drug A group was $M = 5.12$, $SD = 4.21$, $n = 250$, while the mean number of panic attacks in the placebo group was $M = 6.87$, $SD = 4.62$, $n = 250$, $t(498) = 4.43$, $p < .001$, $d = .39$. To participate in this trial, your friend would need to go to a local hospital once a week for a 1-hour visit.

Drug B information: In total, 150 participants took Drug B for 3 months, while another 150 participants took a placebo for 3 months in a double-blind study. While taking the "drug," the participants recorded the number of panic attacks they experienced. The mean number of panic

attacks in the Drug B group was $M = 5.47$, $SD = 2.89$, $n = 150$, while the mean number of panic attacks in the placebo group was $M = 8.11$, $SD = 2.94$, $n = 150$, $t(298) = 7.84$, $p < .001$, $d = .70$. To participate in this trial, your friend would need to go to a hospital that is approximately 45 minutes away once a week for 1.5-hour visits.

28. When deciding which of these two drug studies provides more compelling evidence, you should consider their respective sample sizes. At first glance, it is tempting to assume automatically that studies with larger samples are always better. However, that is not the case. As long as a sample size is sufficiently large, an even larger sample is not appreciably better. Compute the standard error of the mean or expected sampling error for both of these studies below. (You will need to compute the pooled *SD* first.)

    The *SEM* for Study A =

    The *SEM* for Study B =

    Based on the *SEM*s for the two studies:

    a. Study A is much more compelling in terms of its sample size.
    b. Study B is much more compelling in terms of its sample size.
    c. Studies A and B are approximately equal when it comes to sample size; both have relatively small *SEM*s, so choosing between these studies should probably be based on something other than sample size.

29. Based on the effect sizes for the two studies:
    a. Study A has a more impressive effect size.
    b. Study B has a more impressive effect size.
    c. Studies A and B have equally impressive effect sizes.

30. Compare the results of the significance test for the two studies.
    a. Study A had sufficient evidence to reject the null.
    b. Study B had sufficient evidence to reject the null.
    c. Both studies had sufficient evidence to reject the null.

31. Compare the relative cost in terms of time commitment (e.g., travel time as well as treatment time).
    a. Drug A requires a greater time commitment.
    b. Drug B requires a greater time commitment.

32. Which drug trial would you suggest to your friend? (You must choose one.) Explain your rationale; explain the relative importance of the respective sample sizes, the results of the hypothesis tests, the effect sizes, and the relative cost in terms of time commitment (e.g., travel time as well as treatment time). You should not just count the number of times that Drug A or B had "better results." You need to talk about which of these four factors is most important when deciding between these two drug trials and conclude your answer with a recommendation.

Activity 10.2: A Two-Tailed Independent *t* Test

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Compute and interpret an independent samples *t* test by hand and using SPSS
- Draw a distribution of *t* scores assuming the null hypothesis is true and a distribution of *t* scores assuming the research hypothesis is true
- Explain the homogeneity of variance assumption and how it is tested
- Explain what the numerator and the denominator of the *t* test each measure
- Explain the effect of using a one-tailed versus a two-tailed test when hypothesis testing
- Explain how it is possible to not reject the null even when it is false

Two-Tailed Independent *t* Example

In a course on reasoning and decision making, you heard about a phenomenon called anchoring and adjustment. The phenomenon suggests that the way a question is asked can influence the answer people give. You want to see if you can replicate this finding. You asked one group of participants, "Is the Nile River longer than 800 miles [1 mile = 1.61 kilometers]?" They answered "yes" or "no." Then you asked them to estimate the Nile's length to the nearest mile. You changed the first question slightly for a second group of participants. You asked them, "Is the Nile longer than 3,000 miles?" Then you asked them to estimate the Nile's length to the nearest mile. The actual length of the Nile is 4,184 miles. You are interested in whether first asking people the 800 or 3,000 "yes–no" question caused the participants to give estimates that were *significantly different*. You know from previous research that these types of estimates tend to form a normal distribution of scores. You decide to use a $\alpha = .05$. Your data are as follows:

Group 1 800 Low Anchor	*Group 2 3,000 High Anchor*
650	600
850	1,000
900	1,999
1,000	2,000
1,000	2,700
1,100	3,000
1,273	4,000
1,320	4,000
2,000	4,300
2,000	5,000
2,000	6,600

1. Match the assumption to the fact that is relevant to that assumption.

   _____ Independence

   _____ Appropriate measurement of the IV and the DV

   _____ Normality

   _____ Homogeneity of variance

   a. Estimates like this tend to form a normal distribution of scores.

   b. Data were collected from one participant at a time.

   c. This assumption will be assessed later by Levene's test.

   d. The IV manipulation is well defined, and the participants' responses were given on an interval/ratio scale.

2. Write H_0 next to the symbolic notations for the null hypothesis and H_1 next to the research hypothesis.

   a. _____ $\mu_{\text{high anchor}} > \mu_{\text{low anchor}}$

   b. _____ $\mu_{\text{high anchor}} \geq \mu_{\text{low anchor}}$

   c. _____ $\mu_{\text{high anchor}} < \mu_{\text{low anchor}}$

   d. _____ $\mu_{\text{high anchor}} \leq \mu_{\text{low anchor}}$

   e. _____ $\mu_{\text{high anchor}} = \mu_{\text{low anchor}}$

   f. _____ $\mu_{\text{high anchor}} \neq \mu_{\text{low anchor}}$

   The first requirement for using SPSS to analyze data is to enter the data correctly. In SPSS, all the information about a given participant goes on a single row. When conducting an independent *t* test, as in this case, you will need two columns. One column should indicate the IV condition each person is in, either the low anchor condition or the high anchor condition. You should use a coding system like "1" for those in the low (800) anchor condition and "2" for those in the high (3,000) anchor condition. So, for the IV column, you would have a column of "1s" and "2s." Create value labels for this coding scheme if you remember how. The second column should be each person's DV, in this case, each participant's estimate of the Nile's length. When you are done entering data, it should look similar to the file shown on in Figure 10.4; it should not look like the two columns of data in the table on page 347.

   To run the independent *t* test:

   - Click on the Analyze menu. Choose Compare Means and then Independent Samples *t* Test.
   - Move the Independent Variable (the one that indicates which group someone is in) into the "Grouping Variable" box and click on "Define." Enter the values you used to designate Group 1 and Group 2 in the appropriate boxes.
   - Move the Dependent Variable (the one that indicates the actual scores of the participants) into the "Test Variables" box. Click on the OK button.

3. What is the standard error of the mean difference?_____

4. Which of the following statements accurately describes what the standard error is measuring in this study? (Note: The standard error is the denominator of the *t* statistic if you calculate it by hand.)
   a. The standard error is the difference between the means of the two groups or conditions.
   b. The standard error is a measure of sampling error; we would expect a mean difference of this size to occur due to sampling error.
   c. The standard error is a measure of sampling error; *with these sample sizes*, we would expect a mean difference of this size to occur due to sampling error.

5. What two values are used to generate the mean difference? (Note: The mean difference is the numerator of the *t* statistic if you calculate it by hand).

6. Give the exact mean difference value for this study. Mean difference = _____.

7. The mean difference (or numerator of the *t* statistic) is the actual difference between the mean of those who were given the 800 anchor and the mean of those who were given the 3,000 anchor. The standard error of the mean difference (or the denominator of the *t* statistic) is the size of the mean difference we would expect by sampling error. Given these two facts, which of the following is true?
   a. If the mean difference is about the same size as the standard error, the null hypothesis is likely to be false.
   b. If the mean difference is about the same size as the standard error, the null hypothesis is likely to be true.

8. What is the assumption of homogeneity of variance? Select all that apply.
   a. The standard deviations for the two sample means are not significantly different from the population standard deviation.
   b. The standard deviations for the two sample means are not significantly different from each other.
   c. The variances for the two populations being compared are equal.
   d. The variances for the two populations being compared are significantly different from each other.

9. Was the homogeneity of variance assumption violated in this study?
   a. Yes, it was violated. The Sig. level for Levene's test was greater than .05.
   b. Yes, it was violated. The Sig. level for Levene's test was less than .05.
   c. No, it was not violated. The Sig. level for Levene's test was greater than .05.
   d. No, it was not violated. The Sig. level for Levene's test was less than .05.

10. What do you do if this assumption is violated?
    a. You cannot do an independent samples *t* if this assumption is violated.
    b. Use the *t*, *df*, and Sig. information from the "Equal variances assumed row."
    c. Use the *t*, *df*, and Sig. information from the "Equal variances not assumed row."

11. How do you use the information in the Sig. column to determine if you should reject or fail to reject the null when doing a two-tailed test?
    a. You divide the Sig. by 2 and then reject the null hypothesis if that number is less than alpha.
    b. You divide the Sig. by 2 and then reject the null hypothesis if that number is greater than alpha.
    c. You reject the null hypothesis if the Sig. is less than alpha.
    d. You reject the null hypothesis if the Sig. is greater than alpha.
    e. You divide the alpha by 2 and then reject the null hypothesis if that number is less than Sig.
    f. You divide the alpha by 2 and then reject the null hypothesis if that number is greater than Sig.

12. SPSS does not compute *d*, and so you need to compute it by hand. To do so, you will need to compute the pooled variance using the *SD* values in the SPSS output.

13. Summarize the results of this study using APA style.

    Those given the 800 anchor gave estimates of the length of the Nile that were significantly lower (M = _____, SD = _____) than those given the 3,000 anchor (M = _____, SD = _____), t(_____) = _____, p = _____ (two-tailed), d = _____.

    (*Note*: The *df*, *t*, and Sig. value should all come from the "Equal variances not assumed row.")

14. If we had predicted that the people given 3,000 as an anchor would give *higher* estimates than those given 800 as an anchor, it would have been reasonable to do a one-tailed test. What would the null and research hypotheses be if we had predicted higher estimates from the 3,000 group? Write H_0 next to the symbolic notations for the null hypothesis and H_1 next to the research hypothesis.
    a. ___ $\mu_{\text{high anchor}} > \mu_{\text{low anchor}}$
    b. ___ $\mu_{\text{high anchor}} \geq \mu_{\text{low anchor}}$
    c. ___ $\mu_{\text{high anchor}} < \mu_{\text{low anchor}}$
    d. ___ $\mu_{\text{high anchor}} \leq \mu_{\text{low anchor}}$
    e. ___ $\mu_{\text{high anchor}} = \mu_{\text{low anchor}}$
    f. ___ $\mu_{\text{high anchor}} \neq \mu_{\text{low anchor}}$

15. Would the *obtained t value* change if we did a one-tailed test?

    Yes No

16. How do you use the Sig. column of the SPSS output to determine if you should reject or fail to reject the null hypothesis when using a *one-tailed* test?

    a. You divide the Sig. by 2 and then reject the null hypothesis if that number is less than the alpha.

    b. You divide the Sig. by 2 and then reject the null hypothesis if that number is greater than the alpha.

    c. You reject the null hypothesis if the Sig. is less than the alpha.

    d. You reject the null hypothesis if the Sig. is greater than the alpha.

    e. You divide the alpha by 2 and then reject the null hypothesis if that number is less than Sig.

    f. You divide the alpha by 2 and then reject the null hypothesis if that number is greater than Sig.

17. In this case, would we reject the null when doing a one-tailed test with an alpha of .01?

    After learning that people tend to anchor on the initial numbers they are given when making estimates, you wonder if this same principle could be used to increase sales in a grocery store. Specifically, you want to know if providing a high anchor for how many of something shoppers should purchase increases the number of things shoppers buy. You enlist the help of two branches of a neighborhood grocery store chain. The stores are in similar locations with similar demographics and with similar sales figures. One store posts signs for a sports drink that advertise the price as 10 for $10 (high anchor). The other store posts signs as 5 for $5 (low anchor). Both signs also indicate that the price for one bottle is $1. The computerized cash registers recorded the number of bottles of the sports drink purchased by each shopper who purchased at least one bottle of the sports drink. The people who saw the low anchor signs purchased an average of 6 bottles ($SD = 2.39$, $n = 15$), while the people who saw the high anchor signs purchased an average of 9.47 bottles ($SD = 3.29$, $n = 15$). The daily sales of sports drinks across a year do form a normal distribution. The number of bottles sold to each shopper who bought any one a given weekend are as follows:

    Group 1: Low anchor (5 for $5): 5, 10, 2, 5, 5, 5, 5, 5, 10, 8, 10, 5, 5, 4, 6

    Group 2: High anchor (10 for $10): 10, 10, 15, 6, 8, 10, 15, 10, 10, 5, 10, 6, 8, 5, 14

18. Match the assumption to the fact that is relevant to that assumption.

    _____ Independence

    _____ Appropriate measurement of the IV and the DV

    _____ Normality

    _____ Homogeneity of variance

    a. The signs in the different stores were different, and the number of bottles sold to each shopper was recorded.

    b. This assumption will be assessed later with Levene's test.

    c. The population of number of sport drinks bought by individual shoppers has a normal shape.

    d. The purchases of individual shoppers was recorded at both stores.

19. Write H_0 next to the null hypothesis and H_1 next to the research hypothesis.

    _____a. $\mu_1 = \mu_2$

    _____b. $\mu_1 \neq \mu_2$

    _____c. $\mu_1 > \mu_2$

    _____d. $\mu_1 < \mu_2$

    _____e. $\mu_1 > \mu_2$

    _____f. $\mu_1 < \mu_2$

20. What is the critical region for this one-tailed test (use $\alpha = .01$)?

    a. $t > 2.4671$

    b. $t < -2.4671$

    c. $t > 2.4671$ or $t < -2.4671$

21. Compute the test statistic (t).

    a. −3.30

    b. 1.05

    c. −3.47

    d. .003

    e. 1.31

22. Should you reject or fail to reject the null hypothesis?

    a. Reject

    b. Fail to reject

23. Compute the effect size (d).

    a. −3.30

    b. 1.05

    c. −3.47

    d. .003

    e. 1.21

24. How large is the effect?

25. Read and evaluate the following APA style summary of the anchoring study at the grocery store. Then determine if there is an error or omission and, if so, identify the problem. (Select all that apply.)

    The low anchor group ($M = 6.00$, $SD = 2.39$) bought significantly less than the high anchor group ($M = 9.47$, $SD = 3.29$), $t(28) = -3.30$, $d = 1.21$.

    a. There are no errors or omissions in the above APA summary.

    b. The means and standard deviations for each condition are missing.

    c. Some of the statistical t test information is missing.

    d. The sentence implies that the means are significantly different when they are not significantly different.

26. Read and evaluate the following APA-style summary of the anchoring study at the grocery store. Then determine if there is an error or omission and, if so, identify the problem. (Select all that apply.)

The low anchor group bought significantly less than the high anchor group, $t(28) = -3.30$, $p = .003$ (one-tailed), $d = 1.21$.

a. There are no errors or omissions in the above APA summary.
b. The means and standard deviations for each condition are missing.
c. Some of the statistical *t* test information is missing.
d. The sentence implies that the means are significantly different when they are not significantly different.

Activity 10.3: How to Choose the Correct Statistic

Learning Objectives

After reading the chapter and completing the homework and this activity, you should be able to do the following:

- Read a research scenario and determine which statistic should be used
- One of the more challenging aspects of this course is choosing the correct statistic for a given scenario. A common strategy is to match the goal described in the scenario with the statistic that accomplishes that goal. Obviously, using this strategy requires you to first identify the research goal. Look at Appendix J to see a decision tree and a table that may help you choose the correct statistic for a given research situation.

Very Basic Example Problems

Determine which statistic should be used in each of the following situations. When you are assessing a given situation, you need to recognize that if a problem does not give you the sample mean (M) or the sample standard deviation (SD), you can always compute these values from the data. However, if a problem does not give you the population mean (μ) or the population standard deviation (σ), you should assume that these values are not known.

1. Do male teachers make more money than female teachers?

2. Do people have less body fat after running for 6 weeks than before they started running?

3. Intelligence quotient (IQ) scores have a population mean of 100 and a standard deviation of 15. Does the college football team have a mean IQ that is significantly greater than 100?

4. Is the mean height of a sample of female volleyball players taller than 68 inches?

More Detailed Example Problems

Determine which statistic should be used in each of the following research scenarios: z for sample mean, single-sample t, related samples t, or independent samples t.

5. Previous studies have shown that exposure to thin models is associated with lower body image among women. A researcher designs a study to determine if very young girls are similarly affected by thin images. Forty kindergarteners are randomly assigned to one of two groups. The first group plays with Barbie dolls for 30 minutes. The second group plays with a doll with proportions similar to the average American woman. After the 30-minute play period, the researcher measures each girl's body image using a graphic rating scale that yields an interval scaled measure of body image. Which statistic should this researcher use to determine if girls who played with Barbie dolls reported lower body image than girls who played with dolls with proportions similar to the average American woman?

6. A teacher of an art appreciation course wants to know if his course actually results in greater appreciation for the arts. On the first day of class, the teacher asks students to complete an art appreciation survey that assesses attitudes toward a variety of forms of art (e.g., painting, theater, sculpture). Scores on the survey range from 1 = *strongly disagree* to 5 = *strongly agree*, and responses to all of the questions are averaged to create one measure of art appreciation. The same survey was given on the last day of class. The teacher analyzes the survey data and finds that scores were significantly higher on the last day of class than on the first day of class. Which statistic should this researcher use to determine if students' art appreciation scores were higher after the class than before the class?

7. An insurance company keeps careful records of how long all patients stay in the hospital. Analysis of these data reveals that the average length of stay in the maternity ward for women who have had a caesarean section is $\mu = 3.9$ days with a standard deviation of $\sigma = 1.2$. A new program has been instituted that provides new parents with at-home care from a midwife for 2 days after the surgery. To determine if this program has any effect on the number of days women stay in the hospital, the insurance company computes the length of stay of a sample of 100 women who participate in the new program and find that their mean length of stay is 3.4 with a standard deviation of 1.4. Which statistic would help determine if the new program is effective at lowering the average length of mothers' hospital stay?

8. Abel and Kruger (2010) recently analyzed the smiles of professional baseball players listed in the Baseball Register. The photos of players were classified as either big smiles or no smiles. The age of death for all players was also recorded. The results revealed that players with big smiles lived longer than those with no smiles. Which statistic could be used to determine if there was a significant difference in life span between those with big versus no smiles?

9. A questionnaire that assesses the degree to which people believe the world is a fair and just place has a mean of $\mu = 50$. A researcher wonders if this belief is affected by exposure to information, suggesting that the world is not a fair and just place. To answer this research question, he conducts a study with 73 students and has them watch a series of videos where bad things happen to good people. After watching these videos, he gives them the questionnaire and finds

that the average score after watching the videos was 48.1 with a standard deviation of 16.2. Which statistic should the researcher use to determine if watching the video significantly reduced endorsement of the view that the world is fair and just?

10. It is well known that acetaminophen reduces physical pain. DeWall et al. (2010) found that the drug can also reduce psychological pain. Another researcher wonders if the same is true of aspirin. To test the efficacy of aspirin in treating psychological pain, they measured participants' psychological pain, gave them the drug, and then again measured their psychological pain. Psychological pain was measured using an interval scale of measurement. Which statistic should be used to determine if aspirin reduced psychological pain?

11. A recent study revealed that the brains of new mothers grow bigger after giving birth. The researchers performed magnetic resonance imaging on the brains of 19 women and found that the volume of the hypothalamus was greater after giving birth than prior to giving birth. Which statistic would researchers use to determine if the volume of the hypothalamus was greater after giving birth than before?

12. A plastic surgeon notices that most of his patients think that plastic surgery will increase their satisfaction with their appearance and, as a result, make them happier. To see if this is actually the case, he asks 52 of his patients to complete a survey 1 week prior to and then again 1 year after the surgery. The survey consisted of 10 questions such as "I am happy" and "Life is good." Responses to each item were scored with 1 = *strongly agree* and 5 = *strongly disagree*. Scores on all survey items summed into one index of happiness. Which statistic should the surgeon use to determine if patients are happier after having plastic surgery than before?

Activity 10.4: Comparing Independent, Matched, and Related Research Designs

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Compute and interpret independent and related samples *t* tests using SPSS
- Explain how the numerator of a *t* could be increased and the consequences
- Explain how the denominator of a *t* could be decreased and the consequences
- Describe the differences between independent samples and related samples designs
- Explain why related samples designs are generally preferable to independent samples designs

Today, we are going to investigate the effect of an IV on a DV using an independent, matched, and related samples design. Specifically, we are going to measure your heart rate while sitting and standing.

1 What effect do you expect people's physical position of sitting versus standing to have on their heart rate? Will these two conditions produce different heart rates? If so, which condition do you expect to have a higher heart rate?

2. We are going to investigate the relationship between physical position and heart rate using three different designs: (1) an independent design, (2) a matched design, and (3) a related design. Our IV for each design will be whether you are sitting or standing, and the DV will be your heart rate. Once these data are collected, we will analyze the data in SPSS and compare the statistical estimates and conclusions generated by each research design.

After we run each *t* test, use the results to complete the following table.

	Independent	*Matched*	*Related*
Number of people in the study (and number of data points)			
df	$(n_1 - 1) + (n_2 - 1)$	$(N - 1)$	$(N - 1)$
Observed mean difference			
Estimate of sampling error			
Obtained *t* value			
Sig. (*p* value)			
Reject or fail to reject (use $\alpha = .05$)			
d			

Compare the estimates of sampling error for each type of design. Which design has the smallest estimate of sampling error? Explain why this type of design has less sampling error than the other two designs. In your explanation, be sure to use the words *individual differences*, *sample size*, and *sampling error*.

3 Compare the obtained *t* values for each type of design. Which design produced the largest obtained *t* value? Explain why this happened. In your explanation, be sure to use the words *individual differences*, *sample size*, and *sampling error*.

4 Which type of design is probably the best choice for studying the effect of physical position on heart rate and why?

5. Write an APA-style summary for all three *t* tests we performed above.

Activity 10.5: Confidence Intervals for Mean Differences Between Independent Samples

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Compute and interpret CIs for means and mean differences between independent samples
- Write complete APA-style statistical reports that include independent *t* tests and CIs
- Identify the distinct purposes of hypothesis testing, effect sizes, and CIs

1. Previously you learned to compute confidence intervals (CIs) for a population mean, a mean difference between a sample mean and a population mean, and a mean difference between two related sample means. In this activity, you will compute a fourth CI for the mean difference between two independent samples. As with all of the previous CIs, you need three numbers to compute the confidence interval. What are the three numbers you need to compute all CIs? (Choose three.)
   a. An *obtained t* value
   b. A point estimate based on sample data
   c. A *critical t* value
   d. An estimate of sampling error (i.e., some kind of *SEM*)
   e. A Type II error rate

2. When the research design involves two independent samples (e.g., a sample of men and a sample of women), the point estimate is the actual difference observed between the independent sample means. Then a margin of error is added to and subtracted from the point estimate to create the CI. What two values are multiplied together to produce a CI's margin of error?
   a. An *obtained t* value
   b. A point estimate based on sample data
   c. A *critical t* value
   d. An estimate of sampling error (i.e., some kind of *SEM*)
   e. A Type II error rate

3. Why would researchers want to know a CI for a mean difference?
   a. CIs reveal the range of ways a study's results might manifest in a population
   b. CIs determine the magnitude of an IV's effect on a DV

INVESTIGATING ANXIETY DISORDERS

4. A clinical psychologist specializing in treating those with anxiety disorders wonders if the average severity of symptoms differs for men and women. If there were a gender difference, it might hint at potential situational or biological factors contributing to the development of anxiety disorders. On the other hand, if anxiety disorders were present equally in men and women, it might suggest other contributing factors. So, her research goal is to determine if men and women *differ* in anxiety severity. Place an H_0 and H_1 in front of the null and research hypotheses, respectively. She wants $\alpha = .05$, two-tailed test.
   a. _____ Women will be higher in anxiety than men.
   b. _____ Men will be higher in anxiety than women.
   c. _____ Men and women will not differ in anxiety.
   d. _____ Men and women will differ in anxiety.

5. She started her investigation by analyzing data from a sample of intake questionnaires her clients completed. She had questionnaires from eight men and eight women. Determine the critical value for this *t* test.
   a. 2.9768
   b. 2.1448
   c. 1.9600
   d. 2.4441

6. The level of anxiety from each person is presented in the following table. Use SPSS to compute the obtained *t* value for this preliminary study. A higher score represents higher anxiety.

Men	*Women*
6	10
8	7
9	8
7	5
10	11
6	9
10	8
7	9

   The obtained *t* value is
   a. −.50
   b. .57
   c. 1.01
   d. .87

7. Should the null hypothesis be rejected?
   a. Yes, the obtained value is less than the critical value.
   b. No, the obtained value is less than the critical value.
   c. Yes, the obtained value is more than the critical value.
   d. No, the obtained value is more than the critical value.

8. Compute the effect size for this study.

9. Compute the 95% CI for the *mean difference* observed in this preliminary study.

   The point estimate is = _____.

   The critical *t* value is = _____.

   The *SEM* is = _____.

   The UB = _____.

   The LB = _____.

10. To compose the APA results summary, you will need the 95% CI for the mean anxiety of men and the 95% CI for the mean anxiety of women. Use the mean for each group as the respective point estimates and the *SD*s and *N*s to compute the respective *SEM*s. You will also need to look up the t_{CI} for each group. Always use the two-tailed .05 *t* value table to find the t_{CI} for a 95% CI.

    Women point estimate = _____

    Women *SD* = _____

    Women *N* = _____

    Women *SEM* = _____

    Women UB = _____

    Women LB = _____

    Men point estimate = _____

    Men *SD* = _____

    Men *N* = _____

    Men *SEM* = _____

    Men UB = _____

    Men LB = _____

11. Below is the clinician's rough draft of a brief summary of the results. Fill in the blanks so that it complies with APA recommendations.

    The mean anxiety level for women ($M = 8.38$, $SD = 1.85$), 95% CI [_____, _____] was not significantly different from that for men ($M = 7.88$, $SD = 1.64$), CI [_____, _____], t(____) = _____, $p < .05$, d = _____, CI [_____, _____].

12. What should this clinician conclude about anxiety levels of men and women?

    a. Based on this study, it looks like the average anxiety levels do not differ for men and women. Although the sample sizes are small, the effect size is also small and so it is very unlikely that the result is a Type II error.

    b. Based on this study, it looks like the average anxiety levels do not differ for men and women. However, the sample size is very small, and it is possible that this result is a Type II error. The study should be replicated with a larger sample.

    c. Based on this study, it looks like the average anxiety levels do not differ for men and women. However, there is quite a bit of sampling error, and it is possible that this result is a Type I error. The study should be replicated with a larger sample.

EVALUATING CELL PHONES' EFFECT ON SLEEP

A public health advocacy organization hires your firm to investigate if cell phones contribute to sleep deprivation in teenagers. Your firm recruits 46 students between the ages of 16 and 18. Half of them are randomly assigned to a cell phone group and are told to use their cell phone for 1 hour right before going to bed. They are allowed to do whatever they like on the phone (i.e., talk, games, etc.) as long as they are using the phone for a full hour right before bed. The other half are told that they can use their cell phone during the day, but they must not look at the phone for the last hour before they go to sleep. Furthermore, students in both groups agree that they will go to bed between 10 and 11 p.m. each night. Every night for 2 weeks, the students wear a sleep monitor that records the number of hours they slept. The average number of hours of sleep obtained over the course of the week was computed for each student. Again, an intern entered the data into an SPSS data file called "CellPhoneSleep.sav." Use SPSS to run the appropriate analyses. By now you know that you will need to produce a significance test, a CI for each mean, an effect size, and a CI for the mean difference. Use a one-tailed test with $\alpha = .05$ and 95% CIs. The instructions for running an independent measures *t* are in the reading for Chapter 10. This analysis will also provide the confidence interval around the mean difference. To obtain the confidence interval around the two sample means (i.e., cell phone and no cell phone), you need to follow these instructions:

- Click on Analyze, Descriptive Statistics, and Explore.
- Move the dependent variable into the Dependent List box.
- Move the independent variable (grouping variable) into the Factor list box.
- Click on the Statistics button.
- In the Explore:Statistics box, select Descriptives and make sure the Confidence Interval for Mean is set at 95%.
- Click on the Continue button and then on the OK button to run the analysis.

13. Now, fill in the blanks in the APA-style report of the results. When reporting numbers in APA style, round to the second decimal place.

    The participants who used a cell phone before bed __(did/did not)___ sleep significantly less (M = ____, SD = ____), 95% CI [____,____] than participants who did not use a cell phone before bed (M = ____, SD = ____), CI [____,____], t(____) = ____, p = ____ (one-tailed), d = ____, CI [____,____].

14. The manager of the public health firm that hired you is surprised that the confidence interval around the mean difference is so large (i.e., over 1 hour) and wants to know what could be done to make that interval smaller. Which of the following would help reduce the width of the interval?

    a. Compute a 99% confidence interval rather than a 95% confidence interval
    b. Redo the study with a larger sample size to reduce sampling error
    c. Do a two-tailed test rather than a one-tailed test

15. The APA publication manual recommends that researchers include confidence intervals, significance tests, and effect sizes. Match the statistical procedure to the distinct information it provides.

_____ Confidence intervals provide . . .

_____ Significance tests provide . . .

_____ Effect sizes provide . . .

a. an index with which to compare how well different treatments work.

b. if the null hypothesis is true, the probability that a result is due sampling error.

c. a range of plausible values for a population parameter.

Chapter 10 Practice Test

To determine if a regular bedtime helps children academically, a researcher obtains a sample of 50 children in kindergarten who have a regular bedtime every night and a separate sample of 50 children who do not have a regular bedtime every night. Each child's scores on a variety of math tests are combined into a single measure of math performance with higher scores indicating better performance. The mean score for the kindergarteners with a regular bedtime was 68.00 (SD = 10.50) while the mean score for the kindergarteners without a regular bedtime was 63.00 (SD = 10.30). The researcher is not sure if a regular bedtime will help with school performance, and so she decides to do a two-tailed test.

1. Match the assumption to the fact that is relevant to that assumption.

   b Independence
   d Appropriate measurement of the IV and the DV
   a Normality
   c Homogeneity of variance

   a. Samples of 30 or more participants increase confidence that this assumption was not violated.
   b. Data were collected from one participant at a time.
   c. This assumption will be assessed later by Levene's test; you can also be confident that this assumption was not violated if the standard deviations for the two groups are similar.
   d. One variable is a grouping variable, and the other variables is an interval/ratio variable.

2. What is the two-tailed research hypothesis?
   a. The children with a regular bedtime will have better math scores than the children without a regular bedtime.
   b. The children with a regular bedtime will have worse math scores than the children without a regular bedtime.
   c. The children with a regular bedtime will have significantly different math scores than the children without a regular bedtime.
   d. The children with a regular bedtime will not have significantly different math scores than the children without a regular bedtime.

3. What is the two-tailed null hypothesis?
   a. The children with a regular bedtime will have better math scores than will the children without a regular bedtime.
   b. The children with a regular bedtime will have worse math scores than will the children without a regular bedtime.
   c. The children with a regular bedtime will have significantly different math scores than will the children without a regular bedtime.
   d. The children with a regular bedtime will not have significantly different math scores than will the children without a regular bedtime.

4. Compute the *df* for this study.
   a. 100
   b. 50
   c. 99
   d. 98
   e. 49

5. Locate the critical value for this *t* test ($\alpha = .05$).
   a. 1.9845
   b. 1.6606
   c. 2.0096
   d. 1.6766

6. Which of the following statements describes the critical region?
   a. Scores that are unlikely if the null hypothesis is true
   b. Sample mean differences that are unlikely if the null hypothesis is true
   c. Scores that are impossible if the null hypothesis is true
   d. *t* statistics that are likely to occur if the research hypothesis is true

7. Compute the *t* statistic.
   a. .48
   b. 5
   c. 2.40
   d. 2.10
   e. 1.86

8. Should the researcher reject the null hypothesis?
   a. Reject
   b. Not reject

9. Compute the effect size (*d*).
   a. .48
   b. 5
   c. 2.40
   d. 2.10
   e. 1.86

10. How large is the effect?
    a. Small
    b. Small to medium
    c. Medium
    d. Medium to large
    e. Large

11. Are the two groups significantly different from each other?
    a. Yes, because the researcher rejected the null hypothesis.
    b. No, because the researcher failed to reject the null hypothesis.
    c. Yes, because the effect size was in the critical region.
    d. No, because the effect size was not in the critical region.

12. Compute the 95% confidence interval around the mean difference.
    a. LB = 1.2, UB = 1.80
    b. LB = −3.10, UB = 6.90
    c. LB = .87, UB = 9.13
    d. LB = 2.92, UB = 7.08

13. Fill in the blanks:

    The children who had a regular bedtime had significantly higher math test scores (M = _____, SD = _____) than children who did not have a regular bedtime (M = _____, SD = _____), t(_____) = _____, $p <$ _____ (two-tailed), d = _____, 95% CI [_____, _____].

    The researcher who did the earlier study on regular bedtimes is not sure why children with regular bedtimes do better on math tests than other children. It is possible that the children with regular bedtimes are somehow different from children without regular bedtimes (e.g., personality differences, parenting differences). To determine if it is the regular bedtime that is important rather than these extraneous factors, the researcher recruits a sample of 50 children who do not currently have a regular bedtime to participate in a study. The children are randomly assigned to two groups. For children in the regular bedtime group, the parents attend a class where they are told about the benefits of a regular bedtime and are given tips about how to implement a regular bedtime at home. For children in the control group, the parents also attend a class, but rather than learning about the benefits of a regular bedtime, they are told about the benefits of reading. Over the course of the next 3 months, the parents are sent several reminders about the information in the classes they attended. At the end of the 3 months, all students took a math test.

14. What is the one-tailed research hypothesis for this study?
    a. $\mu_{\text{regular bedtime}} > \mu_{\text{control}}$
    b. $\mu_{\text{regular bedtime}} \geq \mu_{\text{control}}$
    c. $\mu_{\text{regular bedtime}} < \mu_{\text{control}}$
    d. $\mu_{\text{regular bedtime}} \leq \mu_{\text{control}}$
    e. $\mu_{\text{regular bedtime}} = \mu_{\text{control}}$
    f. $\mu_{\text{regular bedtime}} \neq \mu_{\text{control}}$

15. What is the one-tailed null hypothesis for this study?
    a. $\mu_{\text{regular bedtime}} > \mu_{\text{control}}$
    b. $\mu_{\text{regular bedtime}} \geq \mu_{\text{control}}$
    c. $\mu_{\text{regular bedtime}} < \mu_{\text{control}}$
    d. $\mu_{\text{regular bedtime}} \leq \mu_{\text{control}}$
    e. $\mu_{\text{regular bedtime}} = \mu_{\text{control}}$
    f. $\mu_{\text{regular bedtime}} \neq \mu_{\text{control}}$

16. Use the SPSS output from the independent samples *t* to determine how many children participated in the study.

Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
MathTest	control group	30	63.5667	16.88232	3.08228
	regular sleep group	30	71.4667	18.66166	3.40714

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
MathTest	Equal variances assumed	.883	.351	-1.719	58	.091	-7.90000	4.59445	-17.09680	1.29680
	Equal variances not assumed			-1.719	57.427	.091	-7.90000	4.59445	-17.09875	1.29875

a. 30
b. 60
c. 58

17. Is the homogeneity of variance assumption violated?
a. Yes, the *p* value for Levene's test is less than .05.
b. No, the *p* value for Levene's test is greater than .05.

18. Which line of output should you use to interpret the independent samples *t* test?
a. Equal variances assumed
b. Equal variances not assumed

19. What is the one-tailed *p* value?
a. .351
b. .1755
c. .091
d. .0455

20. Should the researcher reject or fail to reject the null hypothesis?
a. Reject
b. Fail to reject

21. Compute the effect size ($\hat{d}$).
a. .44
b. 1.72
c. 7.90
d. .091

22. Did the regular sleep group receive significantly higher scores on the math test than the control group?
    a. Yes
    b. No

23. Choose the best APA-style summary of these results.
    a. Children assigned to the regular sleep group ($M = 71.47$, $SD = 18.66$) were significantly different from the children assigned to the control group ($M = 63.57$, $SD = 16.88$), $t(58) = -1.72$, $p = .045$ (one-tailed), $d = .44$.
    b. Children assigned to the regular sleep group ($M = 71.47$, $SD = 18.66$) were not significantly different from the children assigned to the control group ($M = 63.57$, $SD = 16.88$), $t(58) = -1.72$, $p = .045$ (one-tailed), $d = .44$.
    c. Children assigned to the regular sleep group had significantly higher scores on the math test ($M = 71.47$, $SD = 18.66$) than children assigned to the control group ($M = 63.57$, $SD = 16.88$), $t(58) = -1.72$, $p = .045$ (one-tailed), $d = .44$.
    d. Children assigned to the regular sleep group did not have significantly higher scores on the math test ($M = 71.47$, $SD = 18.66$) than children assigned to the control group ($M = 63.57$, $SD = 16.88$), $t(58) = -1.72$, $p = .045$ (one-tailed), $d = .44$.

24. For the independent samples t test, what is the numerator?
    a. The observed difference between means
    b. The difference between the means expected due to sampling error
    c. The difference between the null and research hypotheses

25. For the independent samples t test, what is the denominator?
    a. The observed difference between means
    b. The difference between the means expected due to sampling error
    c. The difference between the null and research hypotheses

REFERENCES

Abel, E. L., & Kruger, M. L. (2010). Birth month affects longevity. *Death Studies, 34*(8), 757–763.

Bower, G. H., Karlin, M. B., & Dueck, A. (1975). Comprehension and memory for pictures. *Memory & Cognition, 3*, 216–220.

Cook, D. A., Gelula, M. H., Dupras, D. M., & Schwartz, A. (2007). Instructional methods and cognitive and learning styles in web-based learning: Report of two randomised trials. *Medical Education, 41*(9), 897–905.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge Academic.

DeWall, C. C., MacDonald, G., Webster, G. D., Masten, C. L., Baumeister, R. F., Powell, C., & Eisenberger, N. I. (2010). Acetaminophen reduces social pain: Behavioral and neural evidence. *Psychological Science, 21*(7), 931–937.

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. *Psychological Science in the Public Interest, 9*(3), 105–119.

Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika, 34*(1–2), 28–35.

CHAPTER 11

One-Way Independent Samples ANOVA

LEARNING OBJECTIVES

After reading this chapter, you should be able to do the following:

- Identify when to use an independent samples ANOVA
- Explain the logic of the ANOVA *F* ratio
- Explain how measurement error, individual differences, and treatment effects influence the numerator and the denominator of the *F* ratio
- Write null and research hypotheses using symbols and words
- Complete an ANOVA summary table by computing the degrees of freedom, *SS*s, *MS*s, and *F* ratio
- Define a critical region and determine if you should reject or fail to reject the null hypothesis
- Compute effect sizes and describe them
- Explain when and why post hoc tests are necessary
- Summarize the results of an independent samples ANOVA using American Psychological Association (APA) style
- Use SPSS to compute an independent samples ANOVA, including post hoc tests
- Interpret the SPSS output for an independent samples ANOVA

INDEPENDENT SAMPLES ANOVA

The independent *t* test and the related samples *t* test both compare two sample means to determine if their deviation is more than would be expected by sampling error. Both of these *t* tests share a major limitation in that they can only compare two sample means at a time. An ANOVA is substantially more flexible in that it can compare two *or more* sample means at the same time to determine if the deviation between any pair of sample means is greater than would be expected by sampling error. So a single ANOVA, also known as a one-way ANOVA or a single-factor ANOVA, can compare two treatments (e.g., Treatment A and Treatment B) to each other as well as compare each of these treatments to a control condition (e.g., placebo).

Reading Question

1. A major advantage of the ANOVA is that it can
   a. be done without ever making a Type I error.
   b. be done without ever making a Type II error.
   c. compare the data from more than two conditions with a single analysis.

Other Names

ANOVA is an abbreviation for analysis of variance. An independent ANOVA, as with an independent *t* test, is used when there are different people in each condition of the design. For example, if there were three different treatments that you wanted to compare in an independent samples design, some people would receive Treatment A, others Treatment B, and the remaining Treatment C. This type of design is also referred to as a between-subject design or an independent measures design. A related samples ANOVA, as with a related samples *t* test, is used when the sample means are generated by the same sample or matched samples. Independent samples ANOVAs and related samples ANOVAs that involve just one categorical variable (e.g., one independent variable, IV) are often referred to as one-way ANOVAs. In this book, we only discuss independent samples ANOVAs.

Reading Question

2. Which of the following is another name for an independent samples ANOVA?
   a. Related samples one-way ANOVA
   b. Between-subjects ANOVA

Logic of the ANOVA

All of the significance tests that you have learned thus far (i.e., *z* for a sample mean, single-sample *t* test, independent *t* test, and the related *t* test) share a common logic. Namely, for all of them, you computed the observed difference between two means and then divided this difference by the difference that would be expected due to sampling error. The logic of the ANOVA test is different.

As is probably evident from the name, an analysis of variance analyzes the variance of scores. Specifically, an ANOVA analyzes the variance of scores both *between* and *within* IV conditions in an attempt to determine if the different treatment conditions affect scores differently. For example, you could use an ANOVA to determine if three different treatment conditions lead to different depression scores.

To comprehend the logic of ANOVAs, you must understand that three things affect the variance of scores:

1. *Measurement error:* There will **always** be variance in scores between **people** because variables cannot be measured perfectly.
2. *Individual differences:* There will **always** be variance in scores between **people** because people are naturally different from each other.
3. *Treatment effect:* There **might be** variance in scores between **groups** because groups experienced different IV conditions or treatments.

The first two sources of variance in scores, measurement error and individual differences, will always be present. These two sources of variance are often collectively referred to as *error variance* because they are both components of sampling error. The third source of variance is really what researchers care about. Researchers use the ANOVA to estimate the amount of score variance created by the different IV treatment conditions. By doing the activities in this chapter, you will be able to understand this better. For now, you have enough information to understand the logic of the ANOVA.

Reading Question

3. Which of the following is the best example of variability due to measurement error?
   a. Two people with the same level of depression produce different scores on a depression inventory.
   b. People getting Treatment A have lower depression scores than those getting Treatment B because Treatment A is more effective at helping depressed people.
   c. Some people are more resistant to treatment for depression than others.

Reading Question

4. Which of the following is the best example of variability due to individual differences?
   a. Two people with the same level of depression produce different scores on a depression inventory.
   b. People getting Treatment A have lower depression scores than those getting Treatment B because Treatment A is more effective at helping depressed people.
   c. Some people are more resistant to treatment for depression than others.

Reading Question

5. Which of the following is the best example of variability due to a treatment effect?
   a. Two people with the same level of depression produce different scores on a depression inventory.
   b. People getting Treatment A have lower depression scores than those getting Treatment B because Treatment A is more effective at helping depressed people.
   c. Some people are more resistant to treatment for depression than others.

Reading Question

6. Of the three sources of variance that can potentially influence scores, only two of them always influence scores. Which two of the following sources of variability always influence the variability of scores?
   a. Measurement error
   b. Individual differences
   c. Treatment effect

Reading Question

7. In research studies, researchers are most interested in evaluating which of the following sources of score variance?
   a. Measurement error
   b. Individual differences
   c. Treatment effect

The ANOVA analyzes the *relative* amounts of these three sources of score variability—namely, (1) treatment effects, (2) individual differences, and (3) measurement error—to produce an F statistic. The conceptual formula of the ANOVA is a ratio of the variability *between* treatment conditions to the variability *within* treatment conditions. Two representations of the conceptual formula for an independent ANOVA are shown as follows:

$$F = \frac{\text{Variability between treatment conditions}}{\text{Variability within treatment conditions}}.$$

$$F = \frac{\text{Treatment effect \& individual differences \& measurement error}}{\text{Individual differences \& measurement error}}.$$

The numerator of the *F* ratio is the variability in scores that exists across the different treatment groups (i.e., IV conditions), which is called between-group variability or between-treatment variability. For example, if one condition had high scores and another condition had low scores, there would be a lot of between-group variability. However, if all conditions had similar scores, there would be little between-group variability. There are three possible sources for *between*-treatment variability. It is possible that the different treatments create between-group variability because one treatment is more effective than another (i.e., the treatments affect the dependent variable differently). This is the variability that researchers are most interested in. However, some of the between-group variability is always caused by individual differences and measurement error. Thus, the numerator of the *F* ratio, the between-group variability, consists of *treatment effects, individual differences effects,* and *measurement error*.

Reading Question

8. Which of the following can contribute to between-treatment variance (i.e., the numerator of an *F* ratio)?
   a. Treatment effects
   b. Individual differences
   c. Measurement error
   d. All of the above

The denominator of the *F* ratio is the variability in scores that exists *within* the treatment groups (i.e., IV conditions), which is called within-group variability. There are two possible sources for *within*-treatment variability: (1) individual differences and (2) measurement error. It is important to note that differences in scores *within* each treatment condition are *not* caused by differences in treatment effectiveness because all people *within a particular group* experienced the same treatment. The only sources for differences *within* a treatment condition are *individual differences* and *measurement error*. The denominator of the *F* ratio is often called the *error term* because it only contains variance created by sampling error.

Reading Question

9. Two of the sources of score variance listed below always contribute to within-treatment variance (i.e., the denominator of an *F* ratio). Which source of score variance listed below never contributes to within-treatment variance?
   a. Treatment effects
   b. Individual differences
   c. Measurement error

Reading Question

10. The denominator of the *F* ratio estimates the amount of variability created by
    a. the treatment.
    b. sampling error.

If you look at the sources of variance in the numerator and the denominator of the independent ANOVA, you can see that the only difference is that the numerator includes *treatment effects* and the denominator does not. This fact is critical to the logic of independent ANOVA. Imagine a situation in

which the treatment effect creates zero variance (i.e., the treatment does not work at all). If you replace "treatment effect" in the conceptual formula with a "0," the equation would be as follows:

$$F = \frac{0 \;\&\text{individual differences} \;\&\text{error}}{\text{Individual differences} \;\&\text{error}}.$$

With the treatment effect variance being zero, the F ratio would equal 1 because the numerator and denominator would be the same number. So if the treatment creates no variability in scores, the ANOVA is expected to produce an F statistic value close to 1. Conversely, if the different treatments create a lot of score variability, the F value is expected to be substantially greater than 1. An ANOVA F value *cannot* be negative because it is the ratio of two variances, and variances must be positive.

Reading Question

11. If the null hypothesis is *true,* the ANOVA F value is expected to be close to
    a. 0.
    b. −1.
    c. 1.

Reading Question

12. If the null hypothesis is *false,* the ANOVA F value is expected to be
    a. substantially less than 1.
    b. substantially greater than 1.

AN EXAMPLE ANOVA PROBLEM

You use an independent samples ANOVA to compare the means of two *or more* groups/samples containing different people. For example, suppose you want to compare cognitive behavioral therapy (CBT) and psychodynamic therapy (PDT) as treatments for depression. You identify a sample of people with major depression and randomly divide them into three different groups. One group undergoes CBT for 6 months, a second group undergoes PDT for 6 months, and a third group functions as a control group and receives no treatment (NT). After 6 months, you assess their levels of depression using the Beck Depression Inventory (BDI; scores range from 0–63), with higher scores indicating greater depression. The depression scores you found for each group are listed in Table 11.1. In this study, the IV is the type of treatment (CBT, PDT, or NT), and the DV (dependent variable) is each person's depression score on the BDI.

Table 11.1 Depression Scores After Three Different Types of Treatment

Group 1 *Cognitive Behavioral Therapy*	*Group 2* *Psychodynamic Therapy*	*Group 3* *No Treatment Control*
5	16	14
9	17	19
11	18	16

(Continued)

Table 11.1 (Continued)

Group 1 *Cognitive Behavioral Therapy*	*Group 2* *Psychodynamic Therapy*	*Group 3* *No Treatment Control*
6	13	9
2	10	15
15	19	25
$M_1 = 8.00$ $SD_1 = 4.65$	$M_2 = 15.50$ $SD_2 = 3.39$	$M_3 = 16.333$ $SD_3 = 5.35$

Step 1: Examine Variables to Assess Statistical Assumptions

The statistical assumptions for independent *t* tests and independent ANOVAs are identical. Therefore, you must consider four assumptions. In your study, the depression scores of individuals must be measured without one participant's score influencing another's (*data independence*). Your DV, depression score, must be measured on an interval/ratio scale, and your IV must identify how the three therapeutic treatments are different (*appropriate measurement of variables for independent ANOVA*). The distribution of sample means for each of your conditions must have a normal shape (*normality*). As was the case with some previous examples, the sample sizes in this study (*ns* = 6) are too small to be confident that the distributions of sample means will have a normal shape unless the original populations have a normal shape. You should try to get sample sizes close to 30[1] participants per condition (so, close to 90 participants total in this case). But, for teaching purposes, these smaller sample sizes will work well. Your fourth assumption is *homogeneity of variance*. For ANOVAs, return to the double standard deviation rule; if any one of your conditions has a standard deviation double that of another, the homogeneity of variance assumption might be violated. While there are ways to test for equal variances in ANOVA (see Field, 2013), ANOVAs are generally quite robust to violations of this assumption as long as the sample sizes are approximately equal.

Reading Question

13. You use a one-way independent measures ANOVA when
    a. the IV defines two independent samples and the DV is measured on an interval/ratio scale.
    b. the IV defines two or more independent samples and the DV is measured on an interval/ratio scale.
    c. the IV defines two matched samples and the DV is measured on an interval/ratio scale.
    d. the IV defines one sample, the DV is measured on an interval/ratio scale, and the DV is measured twice on that same sample.

1 With a sample size of 30, it is very likely that the assumption of normality will be met. However, other issues need to be considered when designing a study, including Type I error, statistical power, and effect sizes. When designing a study, we encourage you to consult Jacob Cohen's (1992) classic article.

e. the IV defines one sample and the DV is measured on an interval/ratio scale, and you do not know the population standard deviation.

f. the IV defines one sample and the DV is measured on an interval/ratio scale, and you do know the population standard deviation.

Step 2: State the Null and Research Hypotheses

Your second step is to set up the null and research hypotheses. The null hypothesis always states that all of the populations have the same mean DV scores. In this case, the null states that the three populations of people being studied (i.e., those getting CBT, PDT, or NT) have the same mean depression scores; any differences observed in the samples are due to sampling error. In contrast, the research hypothesis states that the three populations' mean DV scores are *not* the same. In this case, the research hypothesis states that *at least one* of the populations' mean depression scores is significantly different from *at least one* of the others. Note that the research hypothesis is not specifying that *all* of the population means are different, just that *at least one* population mean is different from *at least one* of the others. The research hypothesis is saying that one, or more, of the treatments works better than one, or more, of the others. Table 11.2 summarizes how to write the null and research hypotheses.

Table 11.2 Symbolic and Verbal Representations of Two-Tailed Research and Null Hypotheses for a One-Way Independent Samples ANOVA

	Symbolic	*Verbal*	*Mean Differences Created by*
Research hypothesis (H_1)	$\mu_1 \neq \mu_2$ and/or $\mu_1 \neq \mu_3$ and/or $\mu_2 \neq \mu_3$	*At least one* of the populations' mean depression score is *not equal* to *at least one* of the other populations' mean depression score.	One or more of the treatments being more effective than one or more of the others
Null hypothesis (H_0)	H_0: $\mu_1 = \mu_2 = \mu_3$	The mean depression scores for all populations *are* equal.	Sampling error

Reading Question

14. You use an independent ANOVA when you want to determine if the mean differences among two or more sample means are likely to be due to sampling error and you have
    a. different people in each condition.
    b. the same people in each condition.

Reading Question

15. The null hypothesis for the independent ANOVA is that
    a. all the sample means are different.
    b. all the sample means are the same.
    c. at least one of the sample means is different from at least one of the other sample means.

Reading Question

16. The research hypothesis for the independent ANOVA is that
    a. all the sample means are different.
    b. all the sample means are the same.
    c. at least one of the sample means is different from at least one of the other sample means.

Step 3: Define the Critical Value of *F*

As with *t* tests, you use a critical value to determine if you should reject the null, but you will need two different *df* values to find the *F* critical value. You will need one *df* value based on the number of participants and another *df* value based on the number of groups. If the null hypothesis is true, the obtained *F* value should be close to 1. If the null hypothesis is false, the obtained *F* value should be far greater than 1. But how far from 1 does an *F* value have to be to reject the null hypothesis? If the *F* value is equal to or greater than the critical *F* value, you reject the null hypothesis.

Specifically, you will need the between-treatments degrees of freedom ($df_{between}$) and the within-treatments degrees of freedom ($df_{within(error)}$) to find the critical *F* value. These *dfs* are computed with the following formulas, respectively:

$$df_{between} = g - 1, \text{ where } g \text{ is the number of groups/treatment onditions, and}$$

$$df_{within(error)} = N - g, \text{where } N \text{ is the number of scores in the entire study.}$$

In this case, the $df_{between} = 3 - 1 = 2$, and the $df_{within(error)} = 18 - 3 = 15$. You use these *df* values to find the critical value of *F* in Appendix C. Appendix C contains critical values for alpha levels of .05 and .01 in two separate tables. The $df_{between}$ (in this case, 2) indicates the column, and the $df_{within(error)}$ (in this case, 15) indicates the row to find the *F* critical value. As illustrated in Figure 11.1, the critical value of *F* when the *dfs* are 2 and 15 and using an alpha value of .05 is 3.68. Therefore, if your obtained *F* value is greater than 3.68, you reject the null hypothesis.

Reading Question

17. If the computed *F* value is greater than the critical value of *F*, the null hypothesis should
    a. not be rejected.
    b. be rejected.

Step 4: Computing the Test Statistic (Independent ANOVA)

4a–4d. Completing the ANOVA Summary Table

To compute the *F* ratio, you need to compute two numbers: (1) the variance between-treatment conditions and (2) the variance within-treatment conditions. In this book thus far, you have generally computed the standard deviation to measure variability. However, when doing an ANOVA, you compute the variance as a measure of variability. You may remember that the sample variance is the *mean squared* deviation of the scores from the mean and that it is computed by dividing the *SS* by its *df*. In ANOVA terminology, the variance is referred to as the *mean square* (abbreviated *MS*). Although the terminology is a bit different, the logic is the same. To compute each *MS*, you divide each

Figure 11.1 Finding the Critical Value for an ANOVA With $\alpha = .05$

	df numerator					
df denominator	*1*	*2*	*3*	*4*	*5*	*6*
1	161.45	199.50	215.71	224.58	230.16	233.99
2	18.51	19.00	19.16	19.25	19.30	19.33
3	10.13	9.55	9.28	9.12	9.01	8.94
4	7.71	6.94	6.59	6.39	6.26	6.16
5	6.61	5.79	5.41	5.19	5.05	4.95
6	5.99	5.14	4.76	4.53	4.39	4.28
7	5.59	4.74	4.35	4.12	3.97	3.87
8	5.32	4.46	4.07	3.84	3.69	3.58
9	5.12	4.26	3.86	3.63	3.48	3.37
10	4.96	4.10	3.71	3.48	3.33	3.22
11	4.84	3.98	3.59	3.36	3.20	3.09
12	4.75	3.89	3.49	3.26	3.11	3.00
13	4.67	3.81	3.41	3.18	3.03	2.92
14	4.60	3.74	3.34	3.11	2.96	2.85
15	4.54	3.68	3.29	3.06	2.90	2.79
16	4.49	3.63	3.24	3.01	2.85	2.74

SS by its *df*. These computations are typically done using a software package (e.g., SPSS), and the results are often presented in an ANOVA source table.

An ANOVA source table is a valuable summary of the ANOVA statistical analysis because it shows how the *obtained F* is created. The source table lists the sources of variance discussed in previous sections, between and within treatments (error), in the first column. The column headings indicate steps required to compute an *F*: *SS*, *df*, *MS*, and finally *F*. Table 11.3 is an ANOVA source table. The formulas for *df*, *MS*, *F*, and η^2 are presented because the best way for you to understand the interrelationships among these terms is seeing how each value is generated.

The $SS_{between}$ represents the sum of the variability created by the treatment effect, individual differences, and measurement error. You compute the $SS_{between}$ by computing the *SS* for the three group means and then multiplying this SS_{means} value by *n* (i.e., the number of people in each group). In this case, the three group means are 8.0, 15.5, and 16.3333, respectively. Therefore, your SS_{means} is the *SS* of these three means:

Table 11.3 An ANOVA Source Table With Formulas

	Step 3a: Compute SS	*Step 3b: Compute df*	*Step 3c: Compute MS*	*Step 3d: Compute F*	*Step 4: Compute Effect Size*
Source of Variance	***SS***	***df***	***MS***	***F***	η_p^2
Between treatments	$n(\sum M^2 - \frac{(\sum M)^2}{g})$	$g - 1$	$\frac{SS_{between}}{df_{between}}$	$\frac{MS_{between}}{MS_{error}}$	$\frac{SS_{between}}{SS_{between} + SS_{error}}$
Within treatments (error)	$\Sigma SS_{each\ treatment}$	$N - g$	$\frac{SS_{error}}{df_{error}}$		
Total	$SS_{between} + SS_{error}$	$N - 1$			

Note: g = number of IV conditions ($g = 3$; CBT, PDT, NT), N = number of people in the study (in this case, 18), and n = number of scores in each condition.

Treatment Conditions	*Means (M)*	M^2
CBT	8	64
PDT	15.5	240.25
NT	16.333	266.777
	$\Sigma M = 39.833$	$\Sigma M^2 = 571.027$

$$SS_{Means} = \sum M^2 - \frac{(\sum M)^2}{g} = 571.027 - \frac{(39.833)^2}{3} = 42.129.$$

Therefore, the $SS_{between}$ would be 42.129 multiplied by n:

$$SS_{between} = (SS_M)n = (42.129)6 = 252.78.$$

The above formula for $SS_{between}$ only works when there are an equal number of people in each treatment condition. In situations where the sample sizes are not equal, you would need a different computational formula, but the logic of the $SS_{between}$ is still the same. The resulting value represents the variability created by the different treatment conditions, individual differences, and measurement error. We do not present the computational formula needed when you have unequal sample sizes here. We are more interested in your understanding of the logic of ANOVA and how to interpret its results than in your ability to compute an ANOVA with unequal sample sizes from raw data.

Now that you have your $SS_{between}$, your next step is finding the SS_{error}. The SS_{error} is the sum of the variability created by individual differences and measurement error within each condition. You find the SS_{error} by computing the *SS* for each treatment condition separately and then summing them.

$$\text{The } SS \text{ for CBT is } SS_{CBT} = \sum X^2 - \frac{(\sum X)^2}{n} = 492 - \frac{(48)^2}{6} = 108.$$

$$\text{The } SS \text{ for PDT is } SS_{PDT} = \sum X^2 - \frac{(\sum X)^2}{n} = 1499 - \frac{(93)^2}{6} = 57.5.$$

$$\text{The } SS \text{ for NT is } SS_{NT} = \sum X^2 - \frac{(\sum X)^2}{n} = 1744 - \frac{(98)^2}{6} = 143.333.$$

$$\text{Therefore, the } SS_{error} = SS_{CBT} + SS_{PDT} + SS_{NT} = 308.83.$$

You now need the SS_{total}, which represents the total variability in the entire set of data. You sum the variability between conditions (i.e., $SS_{between}$) and the variability within conditions (i.e., SS_{error}).

$$SS_{total} = SS_{between} + SS_{within(error)} = 252.78 + 308.83 = 561.61.$$

Your final goal is computing the *obtained F*, which is the ratio of $MS_{between}$ over MS_{error}. To compute the *MS*s, you divide each *SS* value by its own *df* value.

$$MS_{between} = \frac{SS_{between}}{df_{between}} = \frac{252.78}{2} = 126.39.$$

$$MS_{within} = \frac{SS_{within(error)}}{df_{within(error)}} = \frac{308.83}{15} = 20.59.$$

Your *obtained F* is computed by dividing the $MS_{between}$ by the $MS_{within(error)}$:

$$F = \frac{MS_{between}}{MS_{within(error)}} = \frac{126.39}{20.59} = 6.14.$$

Table 11.4 is your ANOVA source table for this analysis. Confirm that you understand how each of the *SS*s, *df*s, *MS*s, and *F* values were computed.

Reading Question

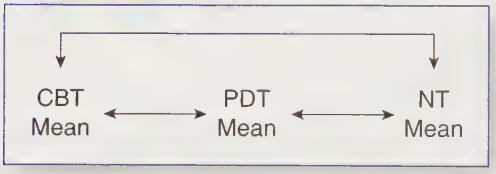
18. The $MS_{between}$ is divided by the $MS_{within(error)}$ to produce the
    a. *SS*.
    b. *df*.
    c. *obtained F* value.

Table 11.4 A Complete ANOVA Source Table

Source of Variance	SS	df	MS	F	η^2
Between treatments	252.78	2	126.39	6.14	.45
Within treatments (error)	308.83	15	20.59		
Total	561.61	17			

In this case, the *obtained F* value was 6.14, which is greater than the critical *F* value of 3.68 ($\alpha = .05$). Therefore, you reject the null hypothesis and conclude that at least one sample mean is significantly different from one of the others.

Figure 11.2 The Three Post Hoc Comparisons in This Study



4e. If Necessary, Compute Post Hoc Tests

The null hypothesis for an ANOVA is that all of the different condition means are equal. Rejecting the null does not indicate which pair(s) of means are different from each other. Thus, if you reject the null hypothesis and there are three or more IV conditions in the study, you need to perform pairwise post hoc tests to determine which pair or pairs of means are different from each other. In this case, three pairwise comparisons are needed. The CBT group needs to be compared with the PDT group and the NT group. In addition, the PDT group needs to be compared with the NT group. Figure 11.2 illustrates the three post hoc comparisons for this analysis.

Reading Question

19. Post hoc tests are needed whenever you
    a. perform an ANOVA.
    b. reject an ANOVA's null hypothesis and there were more than two IV conditions.

There are a number of different post hoc tests that you can use. The one we use throughout this book is Tukey's honestly significant difference (*HSD*) test. The *HSD* test reveals how different any two means must be to be considered "unlikely to result from sampling error" (i.e., significantly different). Therefore, any pair of means that differ by more than the *HSD* value are significantly different. The *HSD* formula is

$$HSD = q\sqrt{\frac{MS_{error}}{n}}.$$

You get the $MS_{within(error)}$ from your overall ANOVA summary table. In this case, $MS_{within(error)} = 20.59$. The number of scores within each treatment condition is n; in this case, $n = 6$. The value of q (called the studentized range statistic) changes, based on the number of treatment conditions (i.e., g), the df for the

error term, and the alpha value. A table of q values is in Appendix D. In this case, $g = 3$, the df for the error term ($df_{within(error)}$) = 15, and the alpha level (α) = .05. Based on these values, the q statistic is 3.67.

	Number of treatment conditions			
df for error term	2	3	4	5
5	3.64 5.70	4.60 6.98	5.22 7.80	5.67 8.42
6	3.46 5.24	4.34 6.33	4.90 7.03	5.30 7.56
7	3.34 4.95	4.16 5.92	4.68 6.54	5.06 7.01
8	3.26 4.75	4.04 5.64	4.53 6.20	4.89 6.62
9	3.20 4.60	3.95 5.43	4.41 5.96	4.76 6.35
10	3.15 4.48	3.88 5.27	4.33 5.77	4.65 6.14
11	3.11 4.39	3.82 5.15	4.26 5.62	4.57 5.97
12	3.08 4.32	3.77 5.05	4.20 5.50	4.51 5.84
13	3.06 4.26	3.73 4.96	4.15 5.40	4.45 5.73
14	3.03 4.21	3.70 4.89	4.11 5.32	4.41 5.63
15	3.01 4.17	3.67 4.84	4.08 5.25	

The top value is the q for $\alpha = .05$
The bottom value is the q for $\alpha = .01$

Thus, your *HSD* value is

$$HSD = q\sqrt{\frac{MS_{error}}{n}} = 3.67\sqrt{\frac{20.59}{6}} = 6.80.$$

Now you compare the *HSD* value to the absolute value of the observed difference between each pair of means. For example, the mean difference between the CBT and PDT conditions is (8 − 15.5) = −7.5. The absolute value of this difference is greater than your *HSD* value of 6.80. Consequently, the CBT group (M = 8) had significantly lower levels of depression than the PDT group (M = 15.5). The two remaining post hoc tests are summarized in Table 11.5. As indicated in the table, significant differences suggest that the difference was not created by sampling error.

Table 11.5 Post Hoc Comparisons

Comparison	*Mean Difference (Absolute Value)*	*Significance*	*Conclusion for Each Comparison*
CBT and PDT	8 – 15.5 = |–7.5|	Significant; more than *HSD* (6.8)	Difference probably due to CBT being a better treatment than PDT
CBT and NT	8 – 16.33 = |–8.33|	Significant; more than *HSD* (6.8)	Difference probably due to CBT being a better treatment than NT
NT and PDT	15.5 – 16.33 = |–.83|	Not significant; less than *HSD* (6.8)	Difference considered to be due to sampling error

Based on these analyses, you can conclude that the CBT group had significantly lower depression scores than the PDT group, suggesting that CBT is better for treating depression than PDT. Similarly, the CBT group had significantly lower depression than the NT group, suggesting that CBT is better than NT. Finally, the PDT and NT groups were not significantly different. A nonsignificant difference between these two conditions means that difference is likely created by sampling error. Therefore, these results suggest that PDT is no better at treating depression than no treatment at all.

The *HSD* formula presented above is appropriate only when the number of scores within each treatment condition is the same. When n is different within each condition, you compute separate *HSD* values for each pairwise comparison. We will use SPSS to compute the *HSD* values when this is the case.

Step 5: Compute the Effect Size and Describe It

When you have two groups, you use d as a measure of effect size. You cannot use d as a measure of effect size for an *overall* ANOVA because there are usually more than two groups, so you cannot compute a simple mean difference. The most common effect size for ANOVAs is partial eta squared (η_p^2). The computations are as follows:

$$\eta_p^2 = \frac{SS_{between}}{SS_{between} + SS_{within(error)}} = \frac{252.78}{252.78 + 308.83} = .45.$$

Table 11.6 General Guidelines for Interpreting η_p^2

η_p^2	*Estimated Size of the Effect*
Close to .01	Small
Close to .06	Medium
Close to .14	Large

For one-way ANOVAs, you should interpret the partial eta squared as the percentage of DV variability the IV "explains." In this case, the type of treatment participants receive "explains" 45% of the variability in their depression scores. In a statistical context, the percentage of variability an IV "explains" is the percentage of DV variance that is associated with changes in the type of treatment participants received. If an IV (i.e., different treatments) explains 45% of the variability in a DV (i.e., depression scores), the remaining 55% of the variability is "explained"

by error. When similar research studies are not available to provide a relative comparison, use the guidelines in Table 11.6 to interpret the size of η_p^2.

The obtained effect size of .45 is greater than .14, and so it is a large effect.

Reading Question

20. The most common effect size used for an ANOVA is η_p^2. If a study were to be done in which the effectiveness of three different treatments for social anxiety were compared and $\eta_p^2 = .22$, which of the following would be the best interpretation of η_p^2?

    a. The anxiety scores resulting from the different treatments were 22% different from each other.

    b. The type of treatment people received explained 22% of the variability in anxiety scores.

You may find that this measure of effect size, partial eta squared, is sometimes called eta squared. While it is true that for one-way ANOVAs, eta squared and partial eta squared yield identical values, they are actually different statistics. The distinction becomes more important when working with the more complex factorial designs described in the next chapter. We use partial eta squared in this book.

Partial eta squared provides an effect size for the overall ANOVA. In the depression example, it tells you that 45% of the variability in depression scores can be explained by the type of treatment participants received. This tells you that the type of treatment had a large effect on depression scores. However, partial eta squared does not describe the size of the effects for the pairwise comparisons. In this case, there are three pairwise comparisons, and you need to compute an effect size for each comparison so you know how much more effective each type of treatment was than the other treatments. For the effect sizes of these pairwise comparisons, you can use the same d you used for the independent t:

$$d = \frac{M_1 - M_2}{\sqrt{SD_p^2}},$$

where the pooled variance is

$$SD_p^2 = \frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{(n_1 - 1) + (n_2 - 1)}$$

As a reminder, the means and standard deviations for each group are provided in Table 11.7.

For each pairwise comparison, you compute the mean difference and then divide by the pooled standard deviations for the two groups. The effect sizes for d are interpreted in the same way as in previous chapters (i.e., .2 is small, .5 is medium, and .8 is large).

These pairwise ds indicate that CBT was a lot more effective than PDT (d = 1.84) and NT (d = 1.66). If you were a clinical psychologist trying to find the best way to help your clients, these statistics would be extremely helpful, indicating that CBT was by far the most effective treatment for your clients. After seeing these results, you

Table 11.7 Mean and Standard Deviations for Each Treatment Type

Group	*Mean (Standard Deviation)*
CBT	8.00 (4.65)
PDT	15.50 (3.39)
NT	16.33 (5.35)

Table 11.8 Effect Sizes (d) for Pairwise Comparisons

Comparison	*Mean Difference*	*Pooled Variance*	d
CBT and PDT	8 – 15.5 = –7.50	$SD_p^2 = \frac{(5)4.65^2 + (5)3.39^2}{(5) + (5)} = 16.56$	$d = \frac{-7.50}{\sqrt{16.56}} = -1.84$
CBT and NT	8 – 16.33 = –8.33	$SD_p^2 = \frac{(5)4.65^2 + (5)5.35^2}{(5) + (5)} = 25.12$	$d = \frac{-8.33}{\sqrt{25.12}} = -1.66$
NT and PDT	15.5 – 16.33= –.83	$SD_p^2 = \frac{(5)5.35^2 + (5)3.39^2}{(5) + (5)} = 20.06$	$d = \frac{-.83}{\sqrt{20.06}} = -.19$

should be thinking that CBT is the therapy most supported by the research evidence. But, as mentioned previously, this study's sample size is too small to draw firm conclusions. The preceding study might be considered a pilot study that suggests a larger study is worth doing.

Step 6: Summarize the Results

When reporting an ANOVA, you must report the $df_{between}$, $df_{within(error)}$, the obtained F value, the p value, MSE, η_p^2, and the individual ds for the pairwise comparisons. You must also indicate which treatment conditions were significantly different from each other. The means, standard deviations, and effect sizes for the pairwise comparisons are reported in tables, so they do not need to be repeated in the text. The following is an example of how the above analysis might be summarized in APA format.

An independent ANOVA revealed meaningful differences among the three depression therapies, $F(2, 15) = 6.14$, $p < .05$, $\eta_p^2 = .45$, $MSE = 20.59$. The means and standard deviations for each treatment condition are in Table 11.9. Tukey's *HSD* post hoc tests revealed that the CBT group had lower depression scores than both the PDT group and the no-treatment group. The PDT therapy was no better than receiving no treatment at all. The *HSD* post hoc tests and their effect sizes are in Table 11.10.

Table 11.9 Means and Standard Deviations for Each Condition

Class	*Mean (SD)*
CBT	8.00 (4.65)
PDT	15.50 (3.39)
NT	16.33 (5.35)

AN ADDITIONAL NOTE ON ANOVAS: FAMILY-WISE ERROR AND ALPHA INFLATION

You may be wondering why we bother with ANOVAs if we are just going to do pairwise post hoc tests after we do an ANOVA. Why don't we just do t tests to compare all pairs of means? The answer lies in the error rate for the study (often called family-wise error). When you set the alpha level of a test at .05, you are setting the

Table 11.10 Tukey *HSD* Post Hoc Results and Effect Sizes

Comparison	*Mean Difference*	p	d
CBT vs. PDT	−7.50	.03	−1.84
CBT vs. NT	−8.33	.02	−1.66
PDT vs. NT	−.83	.95	−.19

Type I error rate at 5%. A Type I error is the probability of saying that a treatment worked when it really did *not* work. However, if you do a bunch of *t* tests, *each t* test has a 5% chance of producing a Type I error. If you perform multiple *t* tests, the exact probability that you made *at least one* Type I error is given by the following family-wise error formula shown:

$$\text{Family} - \text{wise error} = 1-(1-\alpha)^{c} \text{ where } c \text{ is the number of comparisons.}$$

If you have just one comparison, the Type I error rate would be

$$1-(1-.05)^{1} = .05, \text{ or } 5\%.$$

However, if you have three comparisons within the same study, the probability that you made *at least one* Type I error is quite a bit higher at .14 or 14%.

$$1-(1-.05)^{3} = .14.$$

Doing a single ANOVA rather than multiple *t* tests keeps the family-wise error rate at .05 for the overall ANOVA. If post hoc tests are needed, there are a variety of tests to choose from, but the general idea behind most of them is that they keep the Type I error rate from escalating.

Reading Question

21. Why is it better to conduct an ANOVA than multiple *t* tests when you are comparing multiple conditions?
    a. An ANOVA provides an *F* value, which is better than multiple *t* values.
    b. A single ANOVA helps keep the probability a Type I error lower than doing multiple *t* tests.

SPSS

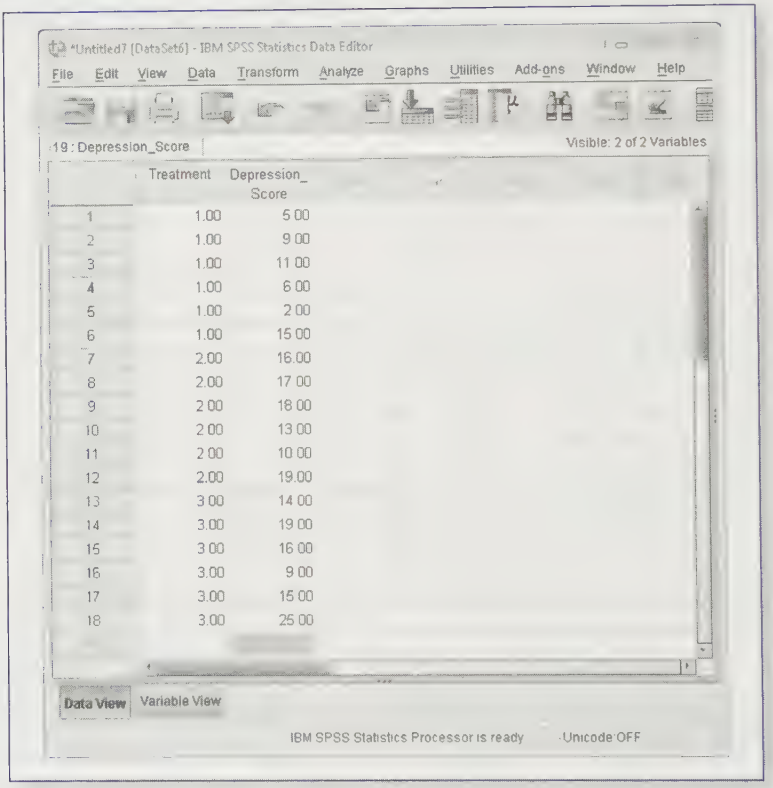
Data File

When creating SPSS data files, enter all the data that came from a single person (or matched persons) on the same row in the data file. For the independent ANOVA, you will have one column to indicate

which treatment each person received and another column to indicate each person's depression score. In the file displayed in Figure 11.1, a "1" indicates that the person received CBT, a "2" indicates PDT, and a "3" indicates that no therapy was given to these individuals. It will help you read the output if you enter this coding system into SPSS. From the "Data View" screen, click on the "Variable View" tab at the bottom left of the screen. On the "Variable View" screen, enter the variable names. Then, across from the treatment variable, click on the cell under the "Values" column heading. When a "button" appears, click on it. In the "Value" box, enter "1." In the "Label" box, enter CBT, and click the "Add" button. Then, enter "2" and PDT in the Value and Label boxes, respectively, and click "Add." Finally, enter "3" and NT, click "Add," and then click "OK."

Your data file should look like the one in Figure 11.3.

Figure 11.3 SPSS Screenshot of Data Entry Screen



To compute an independent samples ANOVA:

- Click on the Analyze menu. Choose General Linear Model and then select Univariate.
- Move the IV into the Fixed Factors box and move the DV into the Dependent Variable box.

- Click on Options and then check Descriptive Statistics and Estimates of Effect Size.
- Click on Continue.
- To obtain post hoc tests, click on the Post Hoc button and then move the IV into the box labeled Post hoc tests for.
- Select the Tukey check box and then Continue.
- Click on OK to run the ANOVA.

Output

For SPSS output, see Figures 11.4 to 11.7.

Reading Question

22. Use the "Descriptive Statistics" output to determine the mean depression score for the patients who received Treatment 2 (PDT). The mean for those who received Treatment 2 was
    a. 8.00.
    b. 15.50.
    c. 16.33.

Reading Question

23. Use the "ANOVA" output to determine what the p value is for the ANOVA test. The p value for the ANOVA test was
    a. 2.
    b. 6.139.
    c. .011.
    d. .04.

Reading Question

24. When you enter data into SPSS to run an independent ANOVA, you should have
    a. a column for the IV (i.e., type of treatment) and a column for the DV (i.e., each person's score).
    b. a column for each DV (i.e., each person's score).

Reading Question

25. Use the "Multiple Comparisons" output to determine which of the following pairs of means are significantly different from each other. Choose all that apply.
    a. Treatments 1 and 2 are significantly different.
    b. Treatments 1 and 3 are significantly different.
    c. Treatments 2 and 3 are significantly different.

Figure 11.4 SPSS Screenshot of Descriptive Statistics

Descriptive Statistics

Dependent Variable:Depression_Score

Treatment	Mean	Std. Deviation	N
1 = CBT	8.0000	4.64758	6
2 = PDT	15.5000	3.39116	6
3 = NT	16.3333	5.35413	6
Total	13.2778	5.74769	18

Mean: Sample means (*M*) for each treatment; total is the overall mean for all scores

Std. Deviation: Sample standard deviations (*s*) for each treatment; total is the overall standard deviation for all scores

***N*:** Sample sizes for each treatment and total number of scores

Figure 11.5 SPSS Screenshot of ANOVA Source Table

Post Hoc Tests

Tests of Between-Subjects Effects

Dependent Variable:Depression_Score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	252.778[a]	2	126.389	6.139	.011	.450
Intercept	3173.389	1	3173.389	154.131	.000	.911
Treatment	252.778	2	126.389	6.139	.011	.450
Error	308.833	15	20.589			
Total	3735.000	18				
Corrected Total	561.611	17				

a. R Squared = .450 (Adjusted R Squared = .377)

Source: Displays the sources of variability in the ANOVA analysis. Only three of the sources are important: **Treatment**, **Error**, and **Corrected Total** **Treatment** + **Error** = **Corrected Total**

Type III Sum of Squares: Displays the *SS* for the **Treatment** (Between Treatment), **Error** (Within Treatments), and **Corrected Total** (Total)

***df*:** Displays the *df* for the **Treatment**, **Error** and **Corrected Total**

Mean Square: Displays the *MS* for the **Treatment** and **Error**

F: *F* value; computed as $\frac{MS_{Treatment}}{MS_{Error}}$

Sig.: The *p* value; reject H_0 if this value (*p* value) is less than α

Partial Eta Squared: this is η_p^2 or effect size

Figure 11.6 An SPSS Screenshot of Post Hoc Tests With Pairwise Comparisons

Homogeneous Subsets

Depression_Score

Tukey *HSD*

(I) Treatment	(J) Treatment	Mean Difference (I–J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1 = CBT	2 = PDT	-7.5000*	2.61973	.030	-14.3047	-.6953
	3 = NT	-8.3333*	2.61973	.016	-15.1380	-1.5287
2 = PDT	1 = CBT	7.5000*	2.61973	.030	.6953	14.3047
	3 = NT	-.8333	2.61973	.946	-7.6380	5.9713
3 = NT	1 = CBT	8.3333*	2.61973	.016	1.5287	15.1380
	2 = PDT	.8333	2.61973	.946	-5.9713	7.6380

Based on observed means.

The error term is mean square (error) = 20.589.

*The mean difference is significant at the .05 level.

Mean Difference (I – J): The difference between the compared treatment means.

The mean difference is significant if a (*) is present

Sig.: The *p* value for each comparison.

The mean difference is significant if this value is less than α.

Redundancies in the Table: Note that all pairwise comparisons are presented twice. For example, the first line is Treatment 1 – Treatment 2 (yielding a difference of –7.5). The third line is Treatment 2 – Treatment 1 (yielding a difference of +7.5).

Figure 11.7 An SPSS Screenshot of Post Hoc Tests With Homogeneous Subsets

Depression_Score

Tukey HSDa

Treatment	N	Subset for alpha = 0.05	
		1	2
1 = CBT	6	8.0000	
2 = PDT	6		15.5000
3 = NT	6		16.3333
Sig.		1.000	.946

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 6.000.

Homogeneous Subsets: This table displays the same information as the previous table in a less precise but more readable way; this table groups means that are not significantly different into the same column; conversely, means that are statistically different are placed in different columns.

OVERVIEW OF THE ACTIVITIES

In Activity 11.1, you will compute an ANOVA as well as post hoc tests by hand. You will also work to understand how within- and between-treatment variability influence the ANOVA. In Activity 11.2, you will work to better understand the different sources of between- and within-treatment variability and how these different sources of variability affect the *F* ratio. In Activity 11.3, you will compute several ANOVAs using SPSS and interpret the output and also summarize the results using an APA-style write-up. You will also compute an *F* and a *t* statistic for the same data set to see how these two statistics are related. In Activity 11.4, you will compute and interpret confidence intervals for independent measures ANOVAs. Finally, in Activity 11.5, you will practice choosing which statistic is appropriate for a given research scenario.

Activity 11.1: Computing One-Way Independent ANOVAs

Learning Objectives

After completing this assignment, you should be able to do the following:

- Describe between-group variance and error variance within the context of independent ANOVA
- Identify researcher choices that would likely decrease error (within-group) variance and increase between-group variance
- Describe the logic of independent ANOVA
- Explain when and why follow-up analyses are needed after rejecting the null hypothesis

Suppose your research team is examining if making more books available to low-income families increases the amount of time those families spend reading to their 2- to 3-year-old children. You identify a sample of 15 low-income families who agreed to participate in a pilot study. Then, you randomly assign 5 families to one of three different experimental conditions. Every family in the first group receives 40 age-appropriate books, every family in the second group receives 20 age-appropriate books, and families in the third group receive 0 books. You train every family to complete a daily journal entry in which they record the number of minutes they spent reading to their children each day. Four months later, you collect the journals and analyze the reported reading times. If making books more available increases family reading times in low-income families, you would expect the greatest reading times in the 40-book group and the lowest reading times in the 0-book group. You chose $\alpha = .05$.

1. The assumptions for an independent ANOVA are the same as those for an independent *t* test. Match the following statistical assumptions to the facts that suggest each assumption is met.

   _____ Independence of data

   _____ Appropriate IV and the DV measurement

   _____ Homogeneity of variance

   _____ Normality

a. The three different conditions are defined and reading times are measured on a ratio scale.
b. The reading times collected from each family do not affect the times collected from other families.
c. The reading times from each group are likely to form a normal distribution.
d. The *SD*s from each group are similar.

2. What is the null hypothesis for this ANOVA?
   a. The population mean reading times for the three groups are not significantly different from each other.
   b. At least one of the population mean reading times for the three groups is significantly different from one of the other population means.
   c. The 40-book families will record significantly more reading time than the 20-book families. The 20-book families will record significantly more reading time than the 0-book families.
   d. The 40-book families will not record significantly more reading time than the 20-book families or the 0-book families.

3. What is the research hypothesis for this ANOVA?
   a. The population mean reading times for the three groups are not significantly different from each other.
   b. At least one of the population mean reading times for the three groups is significantly different from one of the other population means.
   c. The 40-book families will record significantly more reading time than the 20-book families. The 20-book families will record significantly more reading time than the 0-book families.
   d. The 40-book families will not record significantly more reading time than the 20-book families or the 0-book families.

The 15 family reading times (in minutes/day) are listed as follows:

40-Book Families	*20-Book Families*	*0-Book Families*
35	27	24
30	33	29
28	25	22
31	26	25
26	29	20

WITHIN-TREATMENT VARIABILITY (ERROR)

4. The five families in the 40-book condition all reported different reading times, specifically, 35, 30, 28, 31, and 26. These differences in reading times in this one experimental condition generate
   a. within-group variance.
   b. between-group variance.

5. Which of the following might have contributed to the variance in scores of the 40-book families? Choose all that might apply.
   a. The fact that all of these families got 40 books
   b. The fact that the amount of time parents spend at work each day varies greatly between families
   c. The fact that some families are larger than other families
   d. The fact that the ages of the other children in these families (those not between 2 and 3 years old) vary
6. The option(s) you chose for the previous question illustrate(s)
   a. the impact of the treatments on the dependent variable.
   b. the impact of individual differences on the dependent variable.
   c. the impact of measurement error on the dependent variable.
7. The fact that, even after training, few families recorded their exact reading time with complete accuracy illustrates
   a. the impact of the treatments on the dependent variable.
   b. the impact of individual differences on the dependent variable.
   c. the impact of measurement error on the dependent variable.
8. The data from this study also indicate that there was within-group variance in the 20-book group and the 0-book group. Why do these groups have within-group variance?
   a. Because they experienced different treatment conditions than each other
   b. For all the same reasons that the 40-book group had within-group variance
9. The same data are presented below, but this time, the *M*, *SS*, and *SD* are computed for the 40-book and 20-book families. You will have to compute the *M*, *SS*, and *SD* for the 0-book group. (Use your calculator. It will be a lot faster!)

40-Book Families	*20-Book Families*	*0-Book Families*
35	27	24
30	33	29
28	25	22
31	26	25
26	29	20
$M_{40\ books} = 30$ $SS_{40\ books} = 46$ $SD_{40\ books} = 3.39$	$M_{20\ books} = 28$ $SS_{20\ books} = 40$ $SD_{20\ books} = 3.16$	$M_{0\ books} =$ $SS_{0\ books} =$ $SD_{0\ books} =$

Before you move on, confirm that you computed the values for the 0-book group correctly. You should find $M = 24$, $SS = 46$, and $SD = 3.39$.

10. In addition to conceptually understanding the sources of error variability (also called within-treatment variability) in a study, you should also be able to compute the "average" amount of error (within-treatment) variability in a study, which is referred to as the mean square error ($MS_{within(error)}$). To compute the $MS_{within(error)}$, you first compute the $SS_{within(error)}$ and then convert that to a $MS_{within(error)}$. The $SS_{within(error)}$ is relatively easy to find; you simply add the *SS*s for each of the conditions. Look at the above table to find these numbers and add them up. (Remember, each of these *SS* values is the within-group variability for each condition. So, it makes sense that we would add them all together to get the total within-group variability in the entire study.) Find that value now.

    $SS_{within(error)}$ =

11. The $MS_{within(error)}$ is the "average" variability created by individual differences and measurement error. To convert the $SS_{within(error)}$ into the $MS_{within(error)}$, divide the $SS_{within(error)}$ by the $df_{within(error)}$. The $df_{within(error)}$ is the number of people in the entire study (N), minus the number of groups in the study (g). So, $df_{within(error)} = N - g$. Find $df_{within(error)}$ and then use it to create the $MS_{within(error)}$.

    $MS_{within(error)} = SS_{within(error)} / df_{within(error)}$ =

12. Which of the following values represents the "average" amount of variability in the study that is attributable to individual differences and error?
    a. 11
    b. 46.67
    c. 4.24

13. Is the $MS_{within(error)}$ the numerator or the denominator of the *F* ratio?
    a. Numerator
    b. Denominator

14. The $MS_{within(error)}$ represents the amount of variability created by measurement error and:
    a. individual differences
    b. the treatment effect

15. In general, the larger the *F* ratio, the more likely you are to reject the null hypothesis. One way to increase the *F* ratio is to reduce the denominator ($MS_{within(error)}$). Which of the following would reduce the $MS_{within(error)}$? (Choose two)
    a. Increase measurement error
    b. Decrease measurement error
    c. Increase individual differences
    d. Decrease individual differences

16. Match each of the following experimenter actions to its likely consequence.

Experimenter action

____Training participants to record their actual reading times more accurately

____Make sure all parents participating in the study completed high school

____Include participants from many different cultural backgrounds

____Have many different experimenters train participants how to record their reading times

Likely consequence

a. Decrease individual differences

b. Increase individual differences

c. Decrease measurement error

d. Increase measurement error

BETWEEN-TREATMENT VARIABILITY

You just computed the denominator of the *F* ratio, which is the "average" variability within treatment conditions. The numerator of the *F* ratio is the "average" variability between treatment conditions. This next part of the activity focuses on understanding between-treatment variability.

17. The means for the 40-book, 20-book, and 0-book conditions are different (i.e., 30, 28, and 24). These differences in the mean reading times generate

a. within-group variance.

b. between-group variance.

18. Variability between treatments can come from treatment effects, individual differences, and measurement error. Which of the following contribute to the treatment effects?

a. The fact that the families received different numbers of books

b. The fact that the amount of time parents spend at work each day varies greatly between families

c. The fact that some families are larger than other families

d. The fact that the ages of the other children in these families (those not between 2 and 3 years old) vary

e. The fact that few families recorded their exact reading time with complete accuracy

19. Which of the following contribute to individual differences? (Choose all that apply.)

a. The fact that the families received different numbers of books

b. The fact that the amount of time parents spend at work each day varies greatly between families

c. The fact that some families are larger than other families

d. The fact that the ages of the other children in these families (those not between 2 and 3 years old) vary

e. The fact that few families recorded their exact reading time with complete accuracy

20. Which of the following contribute to measurement error?
    a. The fact that the families received different numbers of books
    b. The fact that the amount of time parents spend at work each day varies greatly between families
    c. The fact that some families are larger than other families
    d. The fact that the ages of the other children in these families (those not between 2 and 3 years old) vary
    e. The fact that few families recorded their exact reading time with complete accuracy

21. Which of the following contribute to between-treatments variability but not within-treatments variability?
    a. Treatment effects
    b. Individual differences
    c. Measurement error

22. Which if the following is the numerator of the *F* ratio?
    a. Between-treatments variability
    b. Within-treatments variability

23. How is the numerator of the *F* ratio different from the denominator of the *F* ratio?
    a. The numerator includes treatment effects, individual differences, and measurement error. The denominator includes only individual differences and measurement error.
    b. The denominator includes treatment effects, individual differences, and measurement error. The numerator includes only individual differences and measurement error.

24. If the treatment has an effect on the dependent variable, which should be larger, the numerator or the denominator?
    a. Numerator
    b. Denominator

25. Between-group variance is created when the values in different conditions differ from each other. Between-group variance is harder to see because you are comparing all of the numbers in each condition to all of the numbers in the other conditions. The way to do this is to first compute the *SS* for the means.

Treatment Conditions	*Means (M)*	M^2
40 books	30	900
20 books	28	784
0 books	24	576
	$\Sigma M =$ _____	$\Sigma M^2 =$ _____

$$SS_{means} = \Sigma M^2 - \frac{(\Sigma M)^2}{g} =$$

26. The SS_{means} you just computed represents the amount of variability among the three condition means, but we need to adjust this value slightly to account for the sample size in each condition. As you hopefully recall, sample size is important to consider when evaluating variability. You can convert this SS_{means} into the $SS_{between}$ (a sum of all the between-group variability, which is adjusted for sample size) by multiplying the SS_{means} by n, the number of participants in each condition. Do that below.

$SS_{between} = SS_{means}\ (n) =$

27. The next step is to convert the $SS_{between}$, the sum of all between-group variability, into the "average" or mean between-group variability, called the $MS_{between}$. This is done by dividing the $SS_{between}$ by the number of groups in the study $(g) - 1$, called the $df_{between}$. Compute the $df_{between}$ and then convert the $SS_{between}$ into the "average" between-group variability by using the following equation:

$MS_{between} = SS_{between}/df_{between} =$

F RATIO

28. The $MS_{between}$ value represents the "average" variability created by treatment differences, individual differences, and measurement error. You can now use this value to create the F ratio we need to determine if the three treatment conditions are differentially effective. Compute the F ratio below.

$F = MS_{between}/MS_{within(error)} =$

29. The value you just computed is the *obtained F value*. It is literally the ratio of between-group variance to within-group (error) variance created by the study.

$$F = \frac{MS_{between}}{MS_{within(error)}}.$$

$$F = \frac{\text{treatment effect \& individual differences \& measurement error}}{\text{individual differences + measurement error}}.$$

In this case, this F ratio indicates that there is _____ times more between-group variance than there is within-group variance.

30. Now you need the critical value of F to determine if this ratio is large enough to conclude that the different treatments are differentially effective. Look up the critical F value in Appendix C. The $df_{between}$ indicates what column to look in and the $df_{within(error)}$ indicates which row to look in. Make sure you are using the correct alpha value of .05 when you find your critical value. What is the critical value for this study?

31. Should you reject or fail to reject the null hypothesis?
    a. Reject
    b. Fail to reject

32. Now that you rejected the null hypothesis, what does that mean?
    a. It indicates that all of the treatments produce sample means that are significantly different from each other. All treatments are differentially effective.

b. It means that at least one (perhaps more than one) of the treatments is more effective than one (perhaps more than one) of the other treatments.

POST HOC TESTS

33. Now that you rejected the null hypothesis, you need to run an analysis to determine which treatment or treatments are more effective. Maybe one treatment is better than all the others or maybe two treatments are equally effective, but both are better than the third. A follow-up analysis will determine which of these possibilities occurred. While there are many different follow-up tests we could use, we will only learn one in this text, the *HSD,* or honestly significant difference follow-up test. This test produces a mean difference that you use to determine if pairs of means are significantly different from each other. Any differences between two means is significant if the difference is larger than the *HSD* value. If two means are different by this amount (or more), the difference is unlikely to have resulted from chance and is probably created by the treatment differences. You compute the *HSD* as $q\sqrt{\frac{MS_{within(error)}}{n}}$. You can look up q in Appendix D by using the number of treatment conditions (g), the $df_{within(error)}$, and the alpha value. Your n is the number of participants in each condition. The $MS_{within(error)}$ comes from the F ratio calculation you did earlier. Compute the *HSD* for this analysis now.

$$HSD = q\sqrt{\frac{MS_{within(error)}}{n}} =$$

34. Compute the differences between each pair of treatment conditions. The first one is done for you.

    40-book vs. 20-book: 30 − 28 = 2

    40-book vs. 0-book:

    20-book vs. 0-book:

35. Based on the *HSD* you computed in Question 33 and the condition mean differences you computed for Question 34, which of the three group mean differences are so large that they are unlikely to have been caused by chance? Choose all that apply.

    a. The difference between the 40-book group and the 0-book group

    b. The difference between the 40-book group and the 20-book group

    c. The difference between the 20-book group and the 0-book group

36. Now that you identified which group differences are larger than the *HSD* you computed in Question 34 (i.e., which ones are *unlikely* to result from sampling error), you need to compute effect sizes for each pairwise comparison (i.e., those you computed in Question 35). This is done with the following formula in which the numerator is the difference between the two conditions being compared and the denominator is the square root of the pooled variance for those two conditions.

$$d = \frac{M_1 - M_2}{\sqrt{SD_p^2}}.$$

where the pooled variance is

$$SD_p^2 = \frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{(n_1 - 1) + (n_2 - 1)}.$$

Compute the effect sizes (d) for each pairwise comparison in the table below. The first two comparisons are completed for you.

Comparison	*Mean Difference*	*Pooled Variance*	d
40-book vs. 20-book	30 – 28 = 2	$SD_p^2 = \frac{(4)3.39^2 + (4)3.16^2}{(4) + (4)} = 10.74$	$d = \frac{2}{\sqrt{10.74}} = .61$
40-book vs. 0-book	30 – 24 = 6	$SD_p^2 = \frac{(4)3.39^2 + (4)3.39^2}{(4) + (4)} = 11.49$	$d = \frac{6}{\sqrt{11.49}} = 1.77$
20-book vs. 0-book			

These effect sizes will help you evaluate the differences between the conditions.

37. Based on the results of this pilot study (significance test, post hocs, and effect sizes), what would you say about the book availability intervention for low-income families?
    a. The intervention seems promising; it is certainly worth doing a full study with much larger sample sizes to determine if these results would be replicated with larger sample sizes.
    b. The intervention results look so promising that you should skip doing the larger study. Just go directly to implementing this intervention immediately. Give everyone 20 books.
    c. The intervention is not worth pursuing. Do not bother doing a larger study.

38. People often report the results of studies like this one in an ANOVA summary table. Fill in the blanks in the summary table below using the values you computed in this activity. You will notice that this table includes a total row for the *SS* and the *df*. To obtain the totals, you can simply add between- and within-treatment values. You can also compute these values directly from the data, but for right now, it is most important that you understand that the between- and within-values add up to the total.

    The final value you need for the table is the effect size (η_p^2). To compute the effect size for the ANOVA, use $\eta_p^2 = SS_{between}/(SS_{between} + SS_{within(error)})$.

Source of Variance	SS	df	MS	F	η_p^2
Between treatments	_____	_____	_____	_____	_____
Within treatments (error)	_____	_____	_____		
Total	_____	_____			

39. In this activity, you did all of the computations by hand. You can obtain *most of* this information using SPSS. To enter the data into SPSS, create one column indicating which group each person was in (40-book, 20-book, or 0-book) and name that variable "Group." The other column should be named "Reading," and it will include the actual time each family recorded. Then, you can enter how many books they were given in the Group column (i.e., 40, 20, or 0) and the reading time in the Reading column.

To run a one-way ANOVA in SPSS, do the following:

- Click on the Analyze menu. Choose General Linear Model and then select Univariate.
- Move the IV into the Fixed Factors box and move the DV into the Dependent Variable box.
- Click on Options and then check Descriptive Statistics and Estimates of Effect Size.
- Click on Continue.
- To obtain post hoc tests, click on the Post Hoc button and then move the IV into the box labeled Post hoc tests for.
- Select the Tukey check box and then Continue.
- Click on OK to run the ANOVA.
- Note: SPSS does not compute the effect sizes of pairwise comparisons, so you will always need to compute these by hand.

Locate each item from the ANOVA summary table you created in Question 38 in the SPSS output. You will notice that SPSS does not give you the *HSD* value for the post hoc tests. Instead, it just tells you which means are significantly different. Look back at the annotated SPSS output in the chapter if you need help interpreting the post hoc tests.

40. Even though this was a small pilot study, your research team wants to demonstrate that it knows what it is doing to potential funding agencies. Therefore, your team generates a formal APA-style report of the results. Complete the APA-style write-up below by filling in the blanks.

A one-way ANOVA with number of books as the independent variable and reading time as the dependent variable revealed a significant effect, $F(1, ___) = _____$, $p = _____$, $MSE = 11.00$, $\eta_p^2 = _____$. The means and the standard deviations for each condition are in Table 1. Reading times were significantly higher amongst families given 40 books compared to _____ books. The difference between 20 and 0 books was not significant, nor was the difference between the 20 and 40. However, all effect sizes were medium to large or large, suggesting that this study should be replicated with a larger sample. These results are in Table 2. Overall, the results of this pilot study suggest that running a larger study is justified.

Table 1. Means and Standard Deviations for Each Condition

Condition	*Mean (SD)*
0 books	24.00 (3.39)
20 books	28.00 (3.16)
40 books	_____(_____)

Table 2. Tukey *HSD* Post Hoc Results and Effect Sizes

Comparison	*Mean difference*	p	d
0 books vs. 20 books	4.00	.18	_____
0 books vs. 40 books	6.00	.04	1.77
20 books vs. 40 books	2.00	_____	.61

Activity 11.2: Computing One-Way ANOVAs in SPSS

Learning Objectives

After completing this assignment, you should be able to do the following:

- Run a one-way ANOVA in SPSS
- Compute effect sizes for pairwise comparisons
- Interpret and write up the results in APA format

In Activity 11.1, you conducted a pilot study assessing the impact of providing books to families on the amount of time spent reading with their children. Armed with your pilot study's results, your research team obtained the necessary funds needed to run a larger study with 101 families in each condition. An intern entered all of the data into an SPSS file called "Book Availability." Now you need to run all of the analyses. First run the one-way ANOVA following these instructions:

- Click on the Analyze menu. Choose General Linear Model and then select Univariate.
- Move the IV into the Fixed Factors box and move the DV into the Dependent Variable box.
- Click on Options and then check Descriptive Statistics and Estimates of Effect Size.
- Click on Continue.
- To obtain post hoc tests, click on the Post Hoc button and then move the IV into the box labeled Post hoc tests for.
- Select the Tukey check box and then Continue.
- Click on OK to run the ANOVA.
- Note: SPSS does not compute the effect sizes of pairwise comparisons, so you will always need to compute these by hand.

1. Fill in the values for the overall ANOVA results:

   $F(____, ____) = _____, p = _____, MSE = _____, \eta_p^2 = _____.$

2. Is the overall ANOVA significant?

3. Based on the post hoc output, which pairwise comparisons were significant? Select all that are significantly different.

a. 40 vs. 0
b. 20 vs. 0
c. 40 vs. 20

4. Now, use the SPSS output you just created to compute the effect sizes for each pairwise comparison. One of the comparisons is provided for you below.

Comparison	*Mean Difference*	*Pooled Variance*	d
40-book vs. 20-book	22.90 − 21.78 = 1.12	$SD_p^2 = \frac{(100)7.90^2 + (100)7.78^2}{(100) + (100)} = 61.47$	$d = \frac{1.12}{\sqrt{61.47}} = .14$
40-book vs. 0-book			
20-book vs. 0-book			

5. Now you have everything you need to write up an APA-style report of your team's results for the larger study. You can use the example under Question 40 as a guide. In addition, we also provided the general format for reporting an ANOVA below. With these two resources, you should be able to compose a report. Some blank tables are provided. At the end of the paragraph, include a sentence in which you make a recommendation based on the results about whether the book availability intervention is worth implementing and which condition you would recommend be implemented.

General format for reporting an ANOVA:

- Tell what kind of ANOVA you ran.
- Tell if the ANOVA found any meaningful (i.e., statistically significant) differences among the group means; include statistical information at the end of this sentence, $F\ (df_{between}, df_{within}) = F$ value, p = Sig., MSE = mean square error, η_p^2 = partial eta squared. If all of the means and standard deviations are in a table, you do not need to include them in this statement. If, however, they are not in a table, you must provide the descriptive statistics in the statement.

Table 1. Means and Standard Deviations for Each Condition

Class	*Mean (SD)*
0 books	
20 books	
40 books	

Table 2. Tukey *HSD* Post Hoc Results and Effect Sizes

Comparison	*Mean Difference*	p	d
0 books vs. 20 books			
0 books vs. 40 books			
20 books vs. 40 books			

- Note: If you rejected the null hypothesis, *at least* one meaningful difference exists among the groups.
- Note: If you did not reject the null, all of the mean differences are small enough to be due to sampling error.

- *If you rejected the null,* you must then tell which mean differences were "significant" (i.e., probably too big to be created by sampling error).

Activity 11.3: Independent ANOVA With SPSS

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following.

- Use SPSS to conduct an independent ANOVA
- Interpret the SPSS output
- Compose an APA-style summary of the ANOVA results
- Describe the relationship between an independent *t* test and an independent ANOVA with two conditions
- Read a research scenario and determine what statistical test is appropriate

COLLEGE STUDENTS' CYNICISM

In this activity, you will work with the data file titled "Cynicism.sav." This file contains data about cynicism toward college from 76 hypothetical students. Each student answered four questions about three different types of cynicism, specifically their cynicism toward the academic, social, and institutional aspects of their college. All questions used the same Likert response format, where 1 = *strongly disagree,* 2 = *disagree,* 3 = *neither agree nor disagree,* 4 = *agree,* and 5 = *strongly agree.* Their responses yielded one average score for each type of cynicism. Scores on the scale form a normal distribution. Example questions for each type of cynicism are listed as follows:

Academic: "For many of my courses, going to class is a waste of time."

Social: "It takes a great deal of effort to find fun things to do here."

Institutional: "I would not recommend this place to anyone."

ACADEMIC CYNICISM

For the first analysis, you are going to conduct an ANOVA to determine if year in school (i.e., freshman, sophomore, junior, senior) is associated with academic cynicism.

1. Circle any statistical assumption that is not met:
   a. Independence
   b. Appropriate measurement of the IV and the DV
   c. Normality
   d. Homogeneity of variance
   e. All assumptions are met for this hypothesis test.

2. Write H_0 next to the verbal description of the null hypothesis and H_1 next to the research hypothesis.
   a. _____The mean level of academic cynicism will be the same for the populations of freshmen, sophomore, juniors, and seniors.
   b. _____At least one mean will be different from the other means.

Running the SPSS ANOVA

Compute a single-factor, independent samples ANOVA using "YearInSchool" as the IV (factor) and "AcademicCynicism" as the DV.

- Click on the Analyze menu. Choose General Linear Model and then select Univariate.
- Move "YearInSchool" into the Fixed Factor box and "Academic" into the Dependent Variable box.
- Click on Options, check Descriptives and Estimates of effect size, and then click Continue.
- Click on Post hoc, and move "YearInSchool" into the Post hoc tests for box.
- Check Tukey, and then click Continue.
- Finally, click OK.

3. Record the mean level of academic cynicism for each year in school below. As always, also include the standard deviation for each group.

Freshmen $M =$ _____, $SD =$ _____

Sophomores $M =$ _____, $SD =$ _____

Juniors $M =$ _____, $SD =$ _____

Seniors $M =$ _____, $SD =$ _____

4. Use the SPSS output to complete the ANOVA summary table.

Source	SS	df	MS	F	p *(Sig.)*	η_p^2
Between treatments	____	____	____	____	____	____
Within treatments (error)	____	____	____			
Total	____	____				

5. What information does the η_p^2 provide in above ANOVA source table?
   a. It is an effect size that indicates the degree to which the variable "year in school" is associated with academic cynicism.
   b. It is an effect size that indicates how large the difference is for each pairwise comparison (e.g., freshmen vs. sophomores or freshmen vs. juniors).
   c. It is a significance test that indicates whether or not there is a significant relationships between "year in school" and academic cynicism.
   d. It is a confidence interval that indicates what the population parameter is likely to be for academic cynicism in the population.

6. Whenever you obtain a significant *F* value and there are more than two means, post hoc tests are required. In this case, we need to compare the freshmen mean, sophomore mean, junior mean, and senior mean for academic cynicism. More specifically, we need to compare all possible pairs of two means. All of the possible pairwise comparisons are listed below. For example, the first line represents the comparison between freshmen and sophomores. Likewise, the second line represents the comparison between freshmen and juniors. The Tukey post hoc tests computed by SPSS indicate which of these comparisons are significantly different from each other. Circle "Yes" if the pairwise comparison is statistically significant and "No" if the comparison is not statistically significant.

Comparison	Is It Significant?
Freshmen and sophomores	Yes/No
Freshmen and juniors	Yes/No
Freshmen and seniors	Yes/No
Sophomores and juniors	Yes/No
Sophomores and seniors	Yes/No
Juniors and seniors	Yes/No

7. You should have found that just one of the pairwise comparisons was significant. Next you need to determine which group had higher levels of academic cynicism. Which group had higher levels of academic cynicism?
   a. Freshmen
   b. Sophomores

8. What was the overall effect size (η_p^2) for this study? (If you asked SPSS to provide an effect size estimate, you should find this value in the output; if not, you can calculate it by hand easily.)

9. What does the overall effect size tell you?
   a. It indicates how much of an effect the variable of year in school had on academic cynicism scores.
   b. It indicates which group of students had the highest academic cynicism scores.

10. Researchers are usually far more interested in the effect size for the pairwise comparisons than the overall effect size for the ANOVA. Unfortunately, SPSS does not compute the effect sizes for

the pairwise comparisons. However, it is not difficult to do these computations by hand using the following formula:

$$d = \frac{M_1 - M_2}{\sqrt{SD_p^2}}.$$

Compute the effect sizes for each pairwise comparison in the following table. The first four comparisons are completed for you.

	Mean Difference	*Pooled Variance*	d
Freshmen vs. sophomores	$2.211 - 2.798 = -.587$	$SD_p^2 = \frac{(18).547^2 + (18).565^2}{(18) + (18)} = .309$	$d = \frac{-.587}{\sqrt{.309}} = -1.06$
Freshmen vs. juniors	$2.211 - 2.675 = -.464$	$SD_p^2 = \frac{(18).547^2 + (18).73^2}{(18) + (18)} = .416$	$d = \frac{-.464}{\sqrt{.416}} = -.72$
Freshmen vs. seniors	$2.211 - 2.474 = -.263$	$SD_p^2 = \frac{(18).547^2 + (18).523^2}{(18) + (18)} = .286$	$d = \frac{-.263}{\sqrt{.286}} = -.49$
Sophomores vs. juniors	$2.798 - 2.675 = .123$	$SD_p^2 = \frac{(18).565^2 + (18).73^2}{(18) + (18)} = .426$	$d = \frac{.123}{\sqrt{.426}} = .19$
Sophomores vs. seniors			
Juniors vs. seniors			

11. What information do the *d*s provide in the above table?
    a. They are effect sizes that indicate the degree to which the variable "year in school" is associated with academic cynicism.
    b. They are effect sizes that indicate how large the difference is for each pairwise comparison (e.g., freshmen vs. sophomores or freshmen vs. juniors).
    c. They are significance tests that indicate whether or not there is a significant relationships between "year in school" and academic cynicism.
    d. They are confidence intervals that indicate what the population parameter is likely to be for academic cynicism in the population.

12. When summarizing the data for an APA-style write-up, you may present the means and standard deviations in one table and the results of the post hoc tests in a separate table. Then, you can verbally describe the pattern of results without having to interrupt the flow of the text with descriptive statistics. Record the means and standard deviations in Table 1.

Table 1 Means and Standard Deviations for Each Class

Group	Mean (*SD*)
Freshmen	_____ (_____)
Sophomores	_____ (_____)
Juniors	_____ (_____)
Seniors	_____ (_____)

13. In Table 2, record the mean differences for each pairwise comparison, the *p* value for the pairwise comparison (from the SPSS output), and the effect size (*d*).

Questions 14–16. Writing ANOVA summary statements may seem complicated because there are quite a few details you need to include. But, if you really understand why you did the ANOVA in the first place, it is not too complicated. The general format for reporting an ANOVA is given below. While there are other acceptable formats, this basic one is a good place to start. Read this format carefully and be sure that you understand it before you proceed.

Table 2 Tukey *HSD* Post Hoc Results With Effect Sizes

Comparison	*Mean Difference*	p	d
Freshmen vs. sophomores	_____	_____	_____
Freshmen vs. juniors	_____	_____	_____
Freshmen vs. seniors	_____	_____	_____
Sophomores vs. juniors	_____	_____	_____
Sophomores vs. seniors	_____	_____	_____
Juniors vs. seniors	_____	_____	_____

General format for reporting an ANOVA:

- Tell what kind of ANOVA you ran.
- Tell if the ANOVA found any meaningful (i.e., statistically significant) differences among the group means; include statistical information at the end of this sentence, $F\,(df_{between}, df_{within}) = F$ value, p = Sig., *MSE* = mean square error, η_p^2 = partial eta squared. If all of the means and standard deviations are in a table, you do not need to include them in this statement. If, however, they are not in a table, you must provide the descriptive statistics in the statement.
    - Note: If you rejected the null hypothesis, *at least* one meaningful difference exists among the groups.
    - Note: If you did not reject the null, all of the mean differences are small enough to be due to sampling error.
- *If you rejected the null,* you must then tell which mean differences were "significant" (i.e., probably too big to be created by sampling error).

Now that you understand the general format for reporting ANOVAs, take on the role of a statistics tutor who is trying to help students write a correct ANOVA summary. Identify what is wrong with each of the following ANOVA summaries. All of them are attempting to report the analysis of academic cynicism.

Student 1

To determine if class standing was associate d with academic cynicism, a one-way, independent samples ANOVA was computed with class standing as the IV and academic cynicism as the DV. This analysis revealed a significant effect.

14. Is anything wrong with Student 1's ANOVA summary? (Choose all that apply.)
    a. It does not identify what kind of ANOVA was used.
    b. It does not indicate if the null hypothesis was rejected or not.
    c. It does not include the needed statistical information, in the correct format.
    d. It does not identify which group differences were meaningful (i.e., statistically significant).
    e. The summary has all of the necessary components.

Student 2

A one-way, independent samples ANOVA revealed that class standing (the IV) had a significant effect on academic cynicism (the DV). Tukey's *HSD* post hoc test indicated that freshmen had significantly lower levels of cynicism than sophomores, and the effect size was large. No other differences were significant. However, several of the effect sizes were medium to large, suggesting that the study should be repeated with a larger sample. Descriptive statistics are provided in Tables 1 and 2.

15. Is anything wrong with Student 2's ANOVA summary? (Choose all that apply)
    a. It does not identify what kind of ANOVA was used.
    b. It does not indicate if the null hypothesis was rejected or not.
    c. It does not include the needed statistical information, in the correct format.
    d. It does not identify which group differences were meaningful (i.e., statistically significant).
    e. The summary has all of the necessary components.

Student 3

A one-way, independent samples ANOVA found a significant relationship between class standing and academic cynicism, $F(3, 72) = 3.88$, $p = .02$, $MSE = .78$, $\eta_p^2 = .14$. The *HSD* follow-up tests found that freshmen were less cynical than sophomores and the effect size was large. None of the other mean differences were significantly different. However, several of the effect sizes were medium to large, so the study should be repeated with a larger sample before drawing firm conclusions. Descriptive statistics are provided in Tables 1 and 2.

16. Is anything wrong with Student 3's ANOVA summary? (Choose all that apply.)
    a. It does not identify what kind of ANOVA was used.
    b. It does not indicate if the null hypothesis was rejected or not.
    c. It does not include the needed statistical information, in the correct format.
    d. It does not identify which group differences were meaningful (i.e., statistically significant).
    e. The summary has all of the necessary components.

SOCIAL CYNICISM

Now, determine if differences in social cynicism among freshmen, sophomores, juniors, and seniors are meaningful (i.e., statistically significant) or likely created by sampling error. The analyses are all included below, but you should run the ANOVA in SPSS to ensure that you understand where the numbers come from.

17. Circle any statistical assumption that is not met:
    a. Independence
    b. Appropriate measurement of the IV and the DV
    c. Normality
    d. Homogeneity of variance
    e. All assumptions are met for this hypothesis test

18. Write H_0 next to the verbal description of the null hypothesis and H_1 next to the research hypothesis.
    a. _____The mean level of social cynicism will be the same for the population of freshmen, sophomore, juniors, and seniors.
    b. _____At least one mean will be different from the other means.

19. Complete the ANOVA summary table.

Source	SS	df	MS	F	p *(Sig.)*	η_p^2
Between treatments	3.67	3	1.22	_____	.17	_____
Within treatments (error)	50.59	72	.70			
Total	54.26	75				

20. Why aren't post hoc analyses necessary for this ANOVA?
    a. The overall ANOVA was not statistically significant.
    b. There were only two groups in the study.
    c. The effect size was less than .05.

21. Because the overall ANOVA was not significant, you know that *none* of the post hoc tests would be significant. Why should you still compute effect sizes even when you fail to reject the null hypothesis?
    a. It is possible that you made a Type I error and the effect size will tell you the probability that you made this type of error.
    b. It is possible that you made a Type II error and the effect size can help you determine if you should repeat the study with a larger sample size.

Questions 22–24. Take on the role of a statistics tutor who is trying to help students write a correct ANOVA summary. Identify what is wrong with each of the following ANOVA summaries. All of them are attempting to report the analysis of social cynicism (Questions 16–20).

Student 1

A one-way, independent ANOVA found that the social cynicism scores were not significantly different across class standing, $F(3, 72) = 1.74$, $p = .17$, $MSE = .70$, $\eta_p^2 = .07$. However, some effect sizes were medium. The study should be repeated with a larger sample. Descriptive statistics are in Tables 1 and 2.

Table 1. Means and Standard Deviations for Each Class

Class	*Mean (SD)*
Freshmen	2.53 (.64)
Sophomores	3.08 (.89)
Juniors	2.97 (.94)
Seniors	3.03 (.87)

Table 2. Tukey *HSD* Post Hoc Results With Effect Sizes

Comparison	*Mean Difference*	p	d
Freshmen vs. sophomores	−.55	.19	−.71
Freshmen vs. juniors	−.45	.36	−.55
Freshmen vs. seniors	−.50	.26	−.62
Sophomores vs. juniors	.11	.98	.12
Sophomores vs. seniors	.05	1.00	.06
Juniors vs. seniors	−.05	1.00	.07

22. Is anything wrong with Student 1's ANOVA summary? (Choose all that apply.)
    a. It does not identify what kind of ANOVA was used.
    b. It does not indicate if the null hypothesis was rejected or not.
    c. It does not include the needed statistical information, in the correct format.
    d. It does not describe the meaningful (i.e., statistically significant) group differences accurately.
    e. The summary has all of the necessary components.

Student 2

The social cynicism scores were not significantly different across class standing, $F(3, 72) = 1.74$, $p = .17$, $MSE = .70$, $\eta_p^2 = .07$. Freshmen were the least cynical, followed by juniors, seniors, and sophomores, respectively. It is worth repeating this study with a larger sample to determine whether these differences are likely the result of sampling error. Descriptive statistics are in Tables 1 and 2.

23. What is wrong with Student 2's ANOVA summary? (Choose all that apply.)
    a. It does not identify what kind of ANOVA was used.
    b. It does not indicate if the null hypothesis was rejected or not.
    c. It does not include the needed statistical information, in the correct format.
    d. It does not describe the group differences accurately.
    e. The summary has all of the necessary components.

Student 3

The one-way, independent ANOVA revealed that social cynicism did not differ significantly across class standing.

24. What is wrong with Student 3's ANOVA summary? (Choose all that apply.)
    a. It does not identify what kind of ANOVA was used.
    b. It does not indicate if the null hypothesis was rejected or not.
    c. It does not include the needed statistical information, in the correct format.
    d. It does not describe the meaningful (i.e., statistically significant) group differences accurately.
    e. The summary has all of the necessary components.

INSTITUTIONAL CYNICISM

Finally, determine if the differences in institutional cynicism are meaningful (i.e., statistically significant) or likely to be created by sampling error. The data for this analysis are provided in the problem. However, you should still run the analyses in SPSS and confirm that you know how to obtain these numbers.

25. Circle any statistical assumption that is not met:
    a. Independence
    b. Appropriate measurement of the IV and the DV
    c. Normality
    d. Homogeneity of variance
    e. All assumptions are met for this hypothesis test.

26. Write H_0 next to the verbal description of the null hypothesis and H_1 next to the research hypothesis.
    a. _____The mean level of institutional cynicism will be the same for the population of freshmen, sophomore, juniors, and seniors.
    b. _____At least one mean will be different from the other means.

27. Complete the following tables by finding the values in the SPSS output.

Source	SS	df	MS	F	p *(Sig.)*	η_p^2
Between treatments	9.10	3	3.03	_____	.012	.139
Within treatments (error)	56.26	72	.78			
Total	65.35	75				

Table 1. Means and Standard Deviations for Each Class

Class	*Mean (SD)*
Freshmen	_____ (_____)
Sophomores	2.25 (1.03)
Juniors	2.28 (.94)
Seniors	2.29 (.91)

Table 2. Tukey *HSD* Post Hoc Results With Effect Sizes

Comparison	*Mean Difference*	p	d
Freshmen vs. sophomores	−.78	.04	−.93
Freshmen vs. juniors	−.80	.03	−1.03
Freshmen vs. seniors	−.82	.03	−1.07
Sophomores vs. juniors	.03	1.00	.03
Sophomores vs. seniors	.04	1.00	.03
Juniors vs. seniors	_____	_____	_____

28. The obtained *F* from an ANOVA indicates
    a. which of the group means are significantly different from each other.
    b. whether or not any of the group means are significantly different from each other.

29. *HSD* post hoc tests indicate
    a. which of the group means are significantly different from each other.
    b. whether or not any of the group means are significantly different from each other.

30. Choose the best APA-style report of the institutional cynicism results.
    a. A one-way, independent ANOVA found that the institutional cynicism scores were not significantly different across class standing, $F(3, 72) = 3.88$, $p = .14$, $MSE = .78$, $\eta_p^2 = .01$. Descriptive statistics are in Tables 1 and 2.
    b. A one-way, independent ANOVA found that the institutional cynicism scores were significantly different across class standing, $F(3, 72) = 3.88$, $p = .01$, $MSE = .78$, $\eta_p^2 = .14$. Freshmen were less cynical than sophomores, sophomores were less cynical than juniors, and juniors were less cynical than seniors. No other differences were significant. Descriptive statistics are in Tables 1 and 2.
    c. A one-way, independent ANOVA found that the institutional cynicism scores were significantly different across class standing, $F(3, 72) = 3.88$, $p = .01$, $MSE = .78$, $\eta_p^2 = .14$. Freshmen were less cynical than sophomores, sophomores were less cynical than juniors, and juniors were less cynical than seniors. Descriptive statistics are in Tables 1 and 2.
    d. A one-way, independent ANOVA found that the institutional cynicism scores were significantly different across class standing, $F(3, 72) = 3.88$, $p = .01$, $MSE = .78$, $\eta_p^2 = .14$. Freshmen were less cynical than sophomores, juniors, and seniors. No other differences were significant. Descriptive statistics are in Tables 1 and 2.

Why Include All That Statistical Information (*dfs*, *F*, *p* Value, *MSE*)?

31. Students sometimes wonder why they have to report all of the statistical information. Why would anyone care about all of these numbers? The reason researchers should report all of this

information is that readers who know what they are doing can re-create the entire ANOVA source table when the statistical information is reported properly. For example, suppose that a researcher reported the following results: $F(3, 56) = 6.40$, $p < .05$, $MSE = 6.25$. You should be able to use this information to "work backward" and complete the following source table using basic algebra. For example, you know that $F = MS_{between}/MS_{within(error)}$ If you know any two of those values, you can solve for the third. The same is true for *MS*. You know that $MS = SS/df$. If you know any two of those values, you can solve for the third.

Source	SS	df	MS	F	η_p^2
Between treatments	_____	_____	_____	_____	_____
Within treatments (error)	_____	_____	_____		
Total	_____	_____			

32. How many treatment conditions (groups) were in this study?
33. How many people participated in this study?
34. What is the critical value for the *F* (using an alpha of .05)?
35. Should you reject or fail to reject the null hypothesis?
36. Why do post hoc tests need to be done?
    a. Need to know which pairwise comparisons are significantly different
    b. Need to know how large of an effect the IV had on the DV
    c. Need to know if a Type I error was made

The next questions you should have are, "OK, so we can re-create the source table. Why is that worth doing? Why would a researcher want to do that?" Well, there are many potential reasons why a researcher might be interested in the details of the source table. One of the most compelling is that if researchers are reading many different experiments on the same topic, they can compare or combine the results across different experiments using a statistical procedure called "meta-analysis." To do this procedure, the researchers need the details that are in the ANOVA source table.

37. For the preceding problems, the data were already entered into SPSS for you. However, you should also be able to enter the data into SPSS. Indicate how you would enter the data from the following independent measures design with three groups into SPSS:

    Group 1: 36, 40, 44

    Group 2: 37, 38, 41

    Group 3: 44, 43, 42

Activity 11.4: Understanding Within- and Between-Group Variability

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Describe within-group and between-group variance within the context of an independent ANOVA
- Identify things that increase and decrease within-group variance and between-group variance
- Explain the logic of the independent ANOVA using the terms *within-group variance* and *between-group variance*
- Explain how a relatively large within-group variance affects the *F* ratio
- Explain how a relatively large between-group variance affects the *F* ratio

INDEPENDENT ANOVA

As you know from the chapter, independent ANOVAs help determine if different levels of an IV affect DV scores differently. It was also mentioned in the reading that ANOVAs compute two types of variance: (1) within-group variability and (2) between-group variability. The following fictional scenario should help you understand within-group variability and between-group variability.

A Fable About Types of Variance, Sunflowers, and Types of Liquid

A few weeks ago, Anton went to a local farmers' market. While there, he talked to a man claiming to hold the record for growing the world's tallest sunflower. Anton asked, "So, how did you do it?" The man said, "The single best thing you can do to grow tall sunflowers is to use cola rather than water." Anton was skeptical. He asked the man if he had any evidence that using cola rather than water actually lead to taller sunflowers. The man said, "Well, I never ran an experiment if that is what you mean." That was the beginning of a now legendary (if fictional) experiment. Anton and the sunflower farmer conducted the following study.

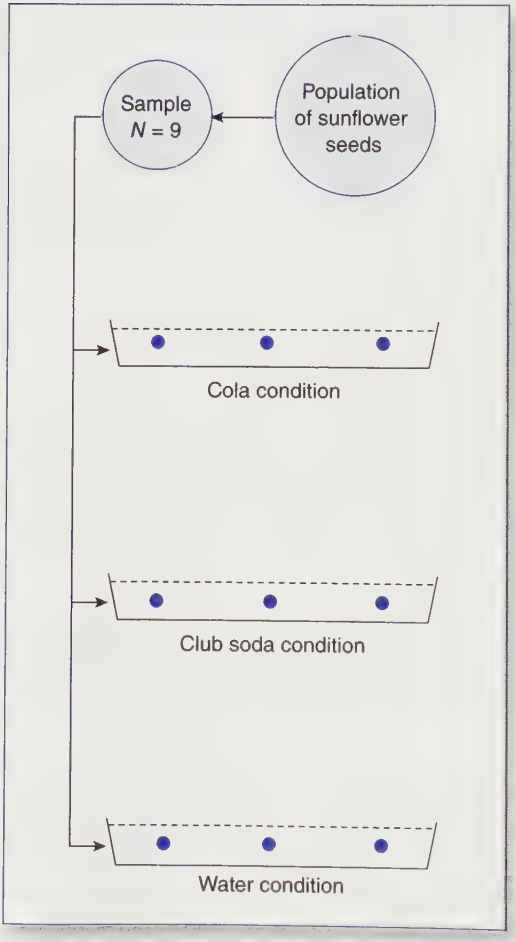
The unlikely colleagues took a sample of sunflower seeds (n = 9) from the farmer's population of seeds. They then found three identical pots and put identical soil in them. They then randomly assigned (planted) 1/3 of their sample in each pot. The figure to the right is a visual representation of the study they designed. The seeds in the first pot were given cola, those in the second pot were given club soda (to see if carbonation was sufficient to increase growth), and those in the third pot were given normal, noncarbonated water.

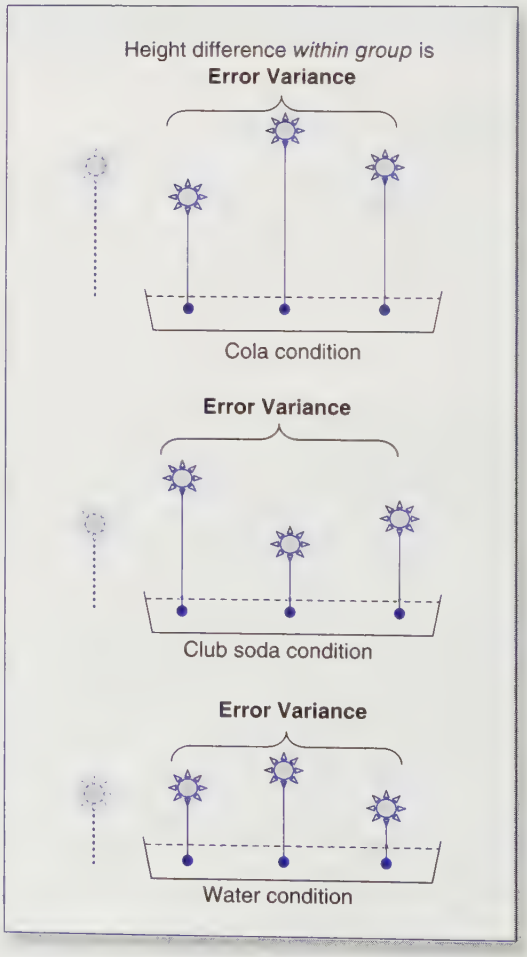
The farmer and Anton took every step possible to ensure that the three pots of seeds got the same amount of sun and the same amount of their designated liquid (i.e., cola, club soda, or water). In short, they tried to make everything identical for the three groups of seeds except for the type of liquid. Logically, then, any *systematic* differences between the heights of the sunflowers in the different conditions would be caused by the different types of liquid the different groups of seeds received (i.e., the IV).

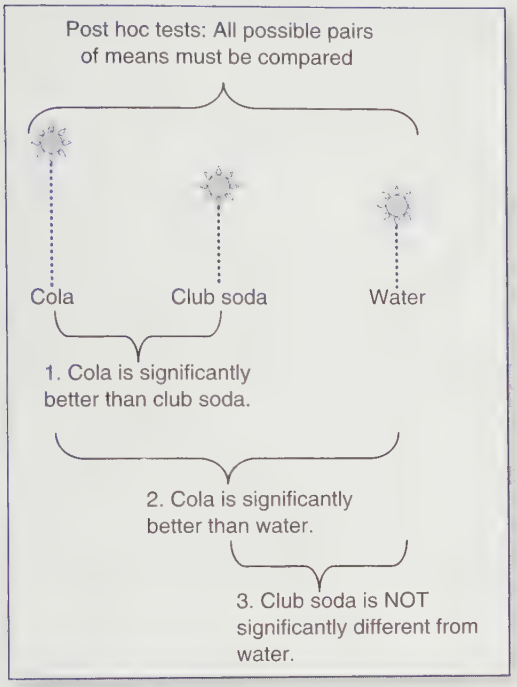
The dependent variable in this study is sunflower height. The figures below represent the height of each sunflower in each IV condition. Clearly, even though sunflowers *within each condition* were treated identically, there is still variability in sunflower height. This variability is called *within-group variance*. Within-group variance is also called error variance because it is created by individual differences in seed quality and/or methodological error (e.g., one seed getting more sun than another). You should notice that there is within-group variability in every condition. These within-group differences are combined to create a single value representing "within-group variability" or "error variability."

Between-group variance is measured by comparing the average sunflower heights produced by each IV condition. The average height produced by each condition is represented by the "dotted sunflowers." Clearly, there is substantial between-group variance in average sunflower height. This between-group variability is created by the different types of liquid in each condition, as well as individual differences and error.

As is explained in the chapter, if the between-group variability is "small" relative to the within-group variability, it is assumed that the IV did not create systematic differences across the different IV conditions. In this case, the between-group variance is relatively large compared to the within-group variance. This would result in a relatively large F value that would lead the researchers to reject the null hypothesis. As a result, the researchers concluded that type of liquid systematically affects sunflower height.







The figure to the left represents the comparison of average sunflower height produced by each IV condition. The researchers have rejected the null hypothesis because the between-group variance (variance created by the IV) was relatively large compared with the within-group variance (variance created by individual differences and error).

Next, researchers perform post hoc tests to determine which pairs of conditions are significantly different from each other. There are three different pairs to compare:

1. Cola versus club soda
2. Cola versus water
3. Club soda versus water

The results of each post hoc comparison are shown to the left.

The cola is significantly better than both club soda and water; furthermore, club soda is *not* significantly different from water.

1. Identify three things that the researchers did that should reduce the within-group variance in this study.
   a. Gave all of the plants the same amount of liquid
   b. Gave some plants cola, some water, and some club soda
   c. Made sure that all of the plants were exposed to the same amount of sunlight
   d. Planted all of the seeds at the same depth

2. Identify the one thing researchers did to cause the between-group variance.
   a. Gave all of the plants the same amount of liquid
   b. Gave some plants cola, some water, and some club soda
   c. Made sure that all of the plants were exposed to the same amount of sunlight
   d. Planted all of the seeds at the same depth

3. Identify two other things that contribute to the between-group variance that the researchers do not want to cause between-group variance.
   a. Type I error
   b. Type II error
   c. Measurement error
   d. Individual differences

4. In the preceding scenario, the researchers rejected the null hypotheses. What is the null hypothesis for an ANOVA comparing three treatment conditions?
   a. H_0: $\mu_1 \neq \mu_2 \neq \mu_3$
   b. H_0: $\mu_1 = \mu_2 = \mu_3$
   c. H_0: $\mu_1 = \mu_2 \neq \mu_3$
   d. H_0: $\mu_1 \neq \mu_2 = \mu_3$

5. In the scenario above, the researchers rejected the null hypotheses. This means that the *F* ratio they computed (i.e., the obtained *F* value) was significantly greater than 1. Why is the *F* ratio greater than 1 if the IV had an effect on the DV?
   a. The variability in the numerator of the *F* is caused by treatment effects, and the variability in the denominator is caused by sampling error. If there is a treatment effect, the sampling error will be equal to zero and the overall *F* will be greater than 1.
   b. The variability in the numerator of the *F* is caused by treatment effects and sampling in error, while the variability in the denominator is only caused by sampling error. If there is a treatment effect, it will lead to a larger numerator compared to the denominator.

Research Team 1		
Group 1: Cola	*Group 2: Club Soda*	*Group 3: Water*
88	83	80
90	85	82
92	87	84
$M_1 = 90.0$	$M_2 = 85.0$	$M_3 = 82.0$
$SD_1 = 2.0$	$SD_2 = 2.0$	$SD_3 = 2.0$
$SS_1 = 8.0$	$SS_2 = 8.0$	$SS_3 = 8.0$

Research Team 2		
Group 1: Cola	*Group 2: Club Soda*	*Group 3: Water*
85	80	77
90	85	82
95	90	87
$M_1 = 90.0$	$M_2 = 85.0$	$M_3 = 82.0$
$SD_1 = 5.0$	$SD_2 = 5.0$	$SD_3 = 5.0$
$SS_1 = 50.0$	$SS_2 = 50.0$	$SS_3 = 50.0$

After reviewing the evidence illustrating the dramatic beneficial effect of "watering" sunflowers with cola, two different research teams attempt to replicate the same study with a different species of sunflowers. As with the previous study you read about, each research team took a sample of nine sunflower seeds from the population of seeds. The seeds were randomly assigned to treatment conditions and their heights were recorded. The data collected from each team are tabulated as follows.

6. If you look at the data within each group, you should be able to see that there is variability within each treatment condition. Flowers within each group were not exactly the same height. For example, for Research Team 1, the heights within the cola condition are all different from each other (88, 90, 92). Identify the two possible reasons for this variability in *within-treatment conditions*.
   a. Treatment effects
   b. Individual differences
   c. Measurement error

7. How can you tell which research team has more error (within-group) variability? It may help you remember that SS_{error} is computed by summing the *SS* for each treatment condition.
   a. Look at the means for the treatment conditions (i.e., cola, club soda, and water). The researcher with larger differences between the means for the three treatment conditions will have more within-treatments variability.
   b. Look at the standard deviation for each treatment condition (i.e., cola, club soda, and water). The researcher with larger standard deviations will have more within-treatments variability.

8. You should remember that the standard deviation is the "standard" distance between each score in a condition and the condition mean. Look at the variability of raw scores within each condition for Research Team 1 and compare it with the variability within each condition for Research Team 2. Which researcher had more within-group variability in their data?
   a. Researcher 1
   b. Researcher 2

9. Which of the following may lead to greater within-treatment variability? Select all that apply.
   a. The researcher is exceptionally careful when determining how much liquid to give each plant and weighs the liquid to the nearest tenth of a gram rather than using a measuring cup.
   b. The lights are fairly dim and so the researcher has a hard time reading the ruler when measuring the sunflowers.
   c. The researcher runs out of one type of sunflower seed, and so he uses a different type of seed from a different supplier.

10. The *SS*s are provided below for each researcher. Use the *SS*s and the information in the problem (i.e., the number of groups and participants) to complete the ANOVA summary tables.

Researcher 1

Source	SS	df	MS	F
Between treatments	98.0	_____	_____	_____
Within treatments	24.0	_____	_____	
Total	122.0	_____		

Researcher 2

Source	SS	df	MS	F
Between treatments	98.0	_____	_____	_____
Within treatments	150.0	_____	_____	
Total	248.0	_____		

11. How do you compute $df_{between}$?
   a. Subtract one from the number of people in the study
   b. Subtract one from the number of treatment conditions
   c. Subtract the number of treatment conditions from the number of people in the study
   d. Subtract the number of people in the study from the number of treatment conditions

12. How do you compute $df_{within(error)}$?
    a. Subtract one from the number of people in the study
    b. Subtract one from the number of treatment conditions
    c. Subtract the number of treatment conditions from the number of people in the study
    d. Subtract the number of people in the study from the number of treatment conditions

13. How do you compute the $MS_{between}$ when you know $SS_{between}$ and $df_{between}$?
    a. Divide the $SS_{between}$ by the $df_{between}$
    b. Multiply the $SS_{between}$ by the $df_{between}$
    c. Divide the $df_{between}$ by the $SS_{between}$

14. Given your answer to the previous question, how do you compute the $MS_{within(error)}$ when you know $SS_{within(error)}$ and $df_{within(error)}$?
    a. Divide the $SS_{within(error)}$ by the $df_{within(error)}$
    b. Multiply the $SS_{within(error)}$ by the $df_{within(error)}$
    c. Divide the $df_{within(error)}$ by the $SS_{within(error)}$

15. After you have the $MS_{between}$ and the $MS_{within(error)}$, how do you compute the obtained F value?
    a. Multiply the $MS_{between}$ times the $MS_{within(error)}$
    b. Divide the $MS_{between}$ by the $MS_{within(error)}$
    c. Divide the $MS_{within(error)}$ by the $MS_{between}$

16. Use the $df_{between}$, the df_{within}, and the table of critical F values in the back of your text to the find the critical value of each of these studies. Use an alpha of .05. What is the critical value for these studies?

    Critical value for Research Team 1 = _____Critical value for Research Team 2 = _____

17. Why are the critical values the same for these two studies? Select all that apply.
    a. Both studies used an alpha of .05, and all F tests use the same critical value when using the same alpha level.
    b. Both studies have the same number of treatment conditions and the same number of sunflowers in each condition.
    c. Both studies have the same $df_{between}$ and $df_{within(error)}$.

18. Should Researcher 1 reject the null hypothesis?
    a. Yes, the obtained F value is less than the critical value.
    b. Yes, the obtained F value is greater than the critical value.
    c. No, the obtained F value is less than the critical value.
    d. No, the obtained F value is greater than the critical value.

19. Should Researcher 2 reject the null hypothesis?
    a. Yes, the obtained F value is less than the critical value.
    b. Yes, the obtained F value is greater than the critical value.
    c. No, the obtained F value is less than the critical value.
    d. No, the obtained F value is greater than the critical value.

20. Based on the differing results of these two studies, explain how two research teams can have identical mean differences (and identical $MS_{between}$) and one reject the null and the other fail to reject the null. What is different about these two studies that made them have different results? Select all that apply.
    a. Researcher 1 had far less within-treatment variability.
    b. Researcher 1 seemed to have less measurement error and/or fewer individual differences than Researcher 2.
    c. Researcher 2 might have been sloppier in conducting the study (e.g., making sure each plant received the same amount of liquid) than Researcher 1.
    d. Researcher 1 had a larger treatment effect than Researcher 2.

After hearing about the success that the farmer had using cola on sunflowers, two other researchers conducted studies to determine if cola, club soda, and water differentially affected the growth of tomato plants. The data are presented as follows.

Researcher 1		
Group 1: Cola	*Group 2: Club Soda*	*Group 3: Water*
49	42	40
52	45	43
55	48	46
$M_1 = 52.0$	$M_2 = 45.0$	$M_3 = 43.0$
$SD_1 = 3.0$	$SD_2 = 3.0$	$SD_3 = 3.0$
$SS_1 = 18.0$	$SS_2 = 18.0$	$SS_3 = 18.0$

Researcher 2		
Group 1: Cola	*Group 2: Club Soda*	*Group 3: Water*
47	44	43
50	47	46
53	50	49
$M_1 = 50.0$	$M_2 = 47.0$	$M_3 = 46.0$
$SD_1 = 3.0$	$SD_2 = 3.0$	$SD_3 = 3.0$
$SS_1 = 18.0$	$SS_2 = 18.0$	$SS_3 = 18.0$

Source	SS	df	MS	F
Between treatments	134.0	_____	_____	_____
Within treatments	54.0	_____	_____	
Total	188.0	_____		

Source	SS	df	MS	F
Between treatments	26.0	_____	_____	_____
Within treatments	54.0	_____	_____	
Total	80.0	_____		

21. Complete the ANOVA summary tables for Researcher 1 and Researcher 2.

22. Look at the standard deviations and *SS*s for each condition in both studies. These values all represent *within-group* variability. Given that all of these values are identical for Researcher 1 and Researcher 2 and that the number of participants is the same in both studies, the _____ for these two studies must also be identical.
    a. $MS_{between}$
    b. MS_{within}

23. You should also notice that the means are different between the treatment conditions of each study. Circle the three possible causes for this between-group variability in this single experiment.
    a. Treatment effects
    b. Effect sizes
    c. Individual differences
    d. Measurement error
    e. Confidence intervals

24. Identify the critical value for each study.

25. Based on these results, explain how two studies can have identical variability in within-treatment conditions (i.e., identical $MS_{within(error)}$) and one study reject the null and the other study fail to reject the null.
    a. The numerators of the *F*s for the two researchers could be different because one researcher has a treatment effect while the other does not.
    b. Even though the denominators are the same, the numerators of the *F*s for the two researchers could be different because one researcher has more measurement error than the other.

The makers of ginger ale are dismayed to hear of the results of these studies because they are sure that their carbonated beverage is every bit as good as cola for "watering" plants. To test the effect of their product on plant growth, they do a study similar to those above but with four different treatment conditions and with nine seeds per condition. The data are as follows.

Ginger Ale	*Cola*	*Club Soda*	*Water*
95	87	80	92
88	90	74	81
87	94	81	90
91	89	78	76
89	98	90	79
83	89	66	78
92	86	93	87
95	87	92	90
84	70	81	68

26. Complete the ANOVA summary table based on the data above. The *SS*s are already provided in the table.

Source	SS	df	MS	F	p *(Sig.)*	η_p^2
Between treatments	399.6	_____	_____	_____	_____	_____
Within treatments	1,757.6	_____	_____			
Total	2,157.2	_____				

27. Based on the results summarized in the above ANOVA summary table, should the null hypothesis for this study be rejected or not?
    a. The null should be rejected because the obtained F value is greater than the critical value.
    b. The null should not be rejected because the obtained F value is less than the critical value.

The following questions ask you to compare the relative amount of within-treatment and/or between-treatment variability that is present in two different data sets. Within-group variability is higher when the scores within the treatment conditions are more different from each other. In other words, studies with more within-group variability have higher *SD*s and *SS*s. Between-group variability is higher when the means for each condition are more different from each other. In other words, studies with more between-group variability have greater differences among the condition means. Use the following two data sets to answer Questions 27 through 31.

Researcher 1		
Group 1: Cola	*Group 2: Club Soda*	*Group 3: Water*
54	59	64
55	60	65
56	61	66
$M_1 = 55.0$	$M_2 = 60.0$	$M_3 = 65.0$
$SD_1 = 1.0$	$SD_2 = 1.0$	$SD_3 = 1.0$
$SS_1 = 2.0$	$SS_2 = 2.0$	$SS_3 = 2.0$

Researcher 2		
Group 1: Cola	*Group 2: Club Soda*	*Group 3: Water*
59	69	79
60	70	80
61	71	81
$M_1 = 60.0$	$M_2 = 70.0$	$M_3 = 80.0$
$SD_1 = 1.0$	$SD_2 = 1.0$	$SD_3 = 1.0$
$SS_1 = 2.0$	$SS_2 = 2.0$	$SS_3 = 2.0$

28. Which researcher's data set has more variability within treatments (i.e., a higher $MS_{within(error)}$)? (Hint: Look at the standard deviations within the treatment conditions for Researcher 1 and Researcher 2.)
    a. Researcher 1
    b. Researcher 2
    c. They have the same amount of within-treatments variability.

29. Which researcher's data set has more variability between treatments (i.e., a higher $MS_{between}$)? (Hint: Look at the means for Researcher 1 and Researcher 2.) Which researcher has more variability between the means?
    a. Researcher 1
    b. Researcher 2
    c. They have the same amount of between-treatments variability.

30. Which researcher's data sets will result in a larger F value? (Hint: The F ratio is computed as $MS_{between}/MS_{within(error)}$.) In this case, the researchers have the same denominator for their F ratio (i.e., the same $MS_{within(error)}$). Which researcher has the larger numerator and hence the larger F?
    a. Researcher 1
    b. Researcher 2
    c. They have the same F value.

31. Which researcher's study is more likely to lead to rejecting the null hypothesis?
    a. Researcher 1
    b. Researcher 2

Use the following two data sets to answer Questions 32 through 36.

Researcher 1		
Group 1: Cola	*Group 2: Club Soda*	*Group 3: Water*
36	36	26
40	38	30
44	42	34
$M_1 = 40$	$M_2 = 38$	$M_3 = 30$
$SD_1 = 4$	$SD_2 = 4$	$SD_3 = 4$
$SS_1 = 32$	$SS_2 = 32$	$SS_3 = 32$

Researcher 2		
Group 1: Cola	*Group 2: Club Soda*	*Group 3: Water*
56	54	46
60	58	50
64	62	54
$M_1 = 60$	$M_2 = 58$	$M_3 = 50$
$SD_1 = 4$	$SD_2 = 4$	$SD_3 = 4$
$SS_1 = 32$	$SS_2 = 32$	$SS_3 = 32$

32. Which researcher's data set has more variability within treatments (i.e., has a higher $MS_{within(error)}$)?
    a. Researcher 1
    b. Researcher 2
    c. They have the same amount of within-treatments variability.

33. Which researcher's data set has more variability between treatments (i.e., has a higher $MS_{between}$)?
    a. Researcher 1
    b. Researcher 2
    c. They have the same amount of between-treatments variability.

34. Which researcher's data sets will result in a larger F value? It is important to note that in this case, both researchers have the same numerator for the F (i.e., $MS_{between}$).
    a. Researcher 1
    b. Researcher 2
    c. They have the same F value.

35. The two studies have the same $df_{between}$ and the same $df_{within(error)}$, which means that they will have the same _____. (Do you know why the dfs for the two studies are the same? If not, figure that out first.)
    a. Obtained F value
    b. Critical value
    c. $MS_{within(error)}$
    d. $MS_{between}$

36. The two studies have the same $MS_{within(error)}$ and the same $MS_{between}$, which means that they will have the same _____. (Do you know why the *MS*s for the two studies are the same, even though all of the scores are different for the two studies? If not, figure that out first.)
    a. Obtained *F* value
    b. Critical value
    c. $MS_{within(error)}$
    d. $MS_{between}$

Use the following two data sets to answer Questions 37 through 41.

Researcher 1		
Group 1: Cola	*Group 2: Club Soda*	*Group 3: Water*
48	58	68
50	60	70
52	62	72
$M_1 = 50$	$M_2 = 60$	$M_3 = 70$
$SD_1 = 2$	$SD_2 = 2$	$SD_3 = 2$
$SS_1 = 8$	$SS_2 = 8$	$SS_3 = 8$

Researcher 2		
Group 1: Cola	*Group 2: Club Soda*	*Group 3: Water*
40	50	60
50	60	70
60	70	80
$M_1 = 50$	$M_2 = 60$	$M_3 = 70$
$SD_1 = 10$	$SD_2 = 10$	$SD_3 = 10$
$SS_1 = 200$	$SS_2 = 200$	$SS_3 = 200$

37. Which researcher's data set has more variability within treatments (i.e., a higher $MS_{within(error)}$)?
    a. Researcher 1
    b. Researcher 2
    c. They have the same amount of within-treatments variability.

38. Which researcher's data set has more variability between treatments (i.e., a higher $MS_{between}$)?
    a. Researcher 1
    b. Researcher 2
    c. They have the same amount of between-treatments variability.

39. Which researcher's data sets will result in a larger *F* value?
    a. Researcher 1
    b. Researcher 2
    c. They have the same amount of within-treatments variability.

40. Which researcher's study is more likely to lead to rejecting the null hypothesis?
    a. Researcher 1
    b. Researcher 2

41. Why is the $MS_{within(error)}$ for Researcher 1 *smaller* than the $MS_{within(error)}$ for Researcher 2?
    a. The scores of Researcher 1 within each treatment condition are more variable (i.e., more different).
    b. The scores of Researcher 1 within each treatment condition are closer together (i.e., more similar).

The following questions are not based on any specific data sets.

42. When a data set has a lot of measurement error and the individuals being measured are very different from each other, the data will tend to have
    a. a large $MS_{within(error)}$.
    b. a large $MS_{between}$.

43. When the different treatments vary in their effectiveness (i.e., some treatments are better than others), the data will tend to have
    a. a large $MS_{within(error)}$.
    b. a large $MS_{between}$.

44. Why do researchers want to minimize within-treatments variability? (Choose two answers.)
    a. Lowering within-treatments variability generally results in higher F values
    b. Lowering within-treatments variability increases the between-treatments variability
    c. Lowering within-treatments variability makes rejecting the null more likely
    d. Lowering within-treatments variability lowers the between-treatments variability

45. Why do researchers want to maximize between-treatments variability? (Choose two answers.)
    a. Increasing between-treatments variability generally results in higher F values
    b. Increasing between-treatments variability decreases the within-treatments variability
    c. Increasing between-treatments variability makes rejecting the null more likely
    d. Increasing between-treatments variability increases the within-treatments variability

46. To determine the effectiveness of a new painkiller on migraines, a medical researcher recruits 80 people with migraines. Twenty of the participants receive a placebo, 20 receive 1 mg of the painkiller, 20 receive 2 mg of the painkiller, and 20 receive 3 mg of the painkiller. Each participant recorded the severity of his or her migraine 1 hour after taking the respective migraine treatments. Which of the following would be the most effective way to increase between-treatments variability in this study?
    a. Use only women in the study
    b. Use fewer participants
    c. Use doses of the drug that are more different (e.g., 10, 20, and 30 mg rather than 1, 2, and 3 mg)

47. Which of the following would be the most effective way to decrease within-treatments variability in this study?
    a. Use only women in the study
    b. Use fewer participants
    c. Give larger doses of the drugs (e.g., 10 mg rather than 1 mg, etc.)

48. This activity focused on explaining within-group and between-group variability of an ANOVA and how each of these types of variability affects the size of the obtained F value. The activity further

emphasized that the larger the obtained F value, the more likely the null hypothesis will be rejected. If an ANOVA is comparing three or more groups, what additional analysis must be performed?

a. Post hoc tests must be performed to determine which of the groups are significantly different from each other.

b. Another ANOVA must be performed.

c. A bunch of t tests must be performed.

49. Finally, try to explain to someone who has never taken a statistics course how an ANOVA compares within-group and between-group variability to determine if three or more treatment options are equally effective.

Activity 11.5: Confidence Intervals

Learning Objectives

- Use SPSS to conduct an independent ANOVA with confidence intervals
- Interpret the SPSS output
- Compose an APA style summary of the ANOVA results, including confidence intervals

In Chapter 8, you learned that confidence intervals help researchers generalize their results to a population. Specifically, they provide a range of plausible values for population parameters. In previous activities, our ANOVA summaries did not include confidence intervals because we wanted you to focus on learning to write ANOVA summaries without the added complexity of confidence intervals. However, the APA publication manual recommends including confidence intervals every time you report a mean or a mean difference (i.e., every time you report the results of t tests or ANOVAs); so, now that you understand how to write ANOVA summaries, it is time to incorporate confidence intervals into your statistical reporting. In this activity, you will compute confidence intervals for ANOVA results and include them in your interpretation of a study.

Agarwal, Karpicke, Kang, Roediger, and McDermott (2007) asked students to read different passages from a textbook. The passages varied in topic (e.g., fossils, wolves, twisters), but all had approximately 1,000 words. Students read each passage once, and then the participants were randomly assigned to use different studying methods. Some students studied with each of the following methods:

1. Studied the material (any way they wanted to) once
2. Studied the material (any way they wanted to) twice
3. Studied the material (any way they wanted to) three times
4. Took a closed-book test and then were given the passage so they could self-grade their answers to the test
5. Took a closed-book test without feedback about the correct answers
6. Took an open-book test
7. Took a test at the same time as they read the passage

One week later, the students took a test over each passage. Overall, they found that test performance was best when students either took a closed-book test with feedback or an open-book test. So, taking some form of test with feedback available led to better learning than studying material three times.

Intrigued by these results, you wonder if these results would generalize to learning about statistics, which you think is quite a bit different than learning about nonstatistics topics. You only have 120 participants, so you decide to test just four experimental conditions: (1) study the material twice, (2) closed-book test, (3) closed-book test with feedback, and (4) open-book test. You randomly assign each of the 120 people to one of the four conditions. All participants got the same statistics chapter to read and studied it according to the method they were assigned. One week later, they took a test on the statistics chapter.

1. Your data are included in the SPSS file titled "TestingEffectANOVA.sav." Run an ANOVA to determine if studying method had an effect on test scores (see Chapter 11, p. 384, for instructions on how to run the ANOVA). Fill in the needed statistical information based on the SPSS output you obtain.

   A one-way independent measures ANOVA revealed that studying method had a significant effect on test scores, F(_____, _____) = _____, $p < .001$, MSE = _____, η_p^2 = _____.

2. Use the SPSS output to record the means and standard deviations for the four groups in the following table:

			95% Confidence Intervals	
	Mean	*Standard Deviation*	*Lower Bound*	*Upper Bound*
a. Studied the material twice				
b. Closed-book test with feedback				
c. Closed-book test without feedback				
d. Open-book test				

3. Compute the 95% confidence intervals:
   a. Click on Analyze→Descriptive→ Explore.
   b. Move the IV (StudyGroup) into the "Factor list" box and the DV (TestScores) into the "Dependent list" box.
   c. Click on the Statistics box and choose 95% confidence intervals.
   d. Click on the Continue button and then the OK button.

Although you can add confidence intervals to APA-style write-ups of ANOVAs, it is often easier for readers to understand if all the numbers are presented in a table format. Thus, it is very common for researchers to present ANOVA results with confidence intervals in a table similar to the one shown above. Add the upper and lower bound values for each confidence interval to the table above.

4. Are the confidence intervals you just computed confidence intervals around means or mean differences?

5. The overall ANOVA was significant and there were more than two groups, so you need to do post hoc tests to determine which groups were significantly different from each other. Based on the SPSS output, which pairwise comparisons are significantly different? If the difference was significant, circle the group with the higher test score.

	Significant?	
Closed-book test without feedback vs. study twice	Yes	No
Closed-book test with feedback vs. study twice	Yes	No
Open-book test vs. study twice	Yes	No
Closed-book test with feedback vs. closed-book test without feedback	Yes	No
Open-book test vs. closed-book test without feedback	Yes	No
Open-book test vs. closed-book test with feedback	Yes	No

6. The post hoc tests allow you to determine which mean differences are statistically significant. You can also compute the confidence intervals around these mean differences to obtain a range of plausible values for these group differences if these experimental conditions were applied to the entire population. SPSS provides the lower and upper bounds for the pairwise mean differences in the Tukey post hoc table. Record the point estimate for each mean difference as well as the lower and upper bounds for each confidence interval in the following table.

		95% Confidence Interval for Mean Difference	
	Mean Difference	*Lower Bound*	*Upper Bound*
Closed-book test without feedback vs. study twice			
Closed-book test with feedback vs. study twice			
Open-book test vs. study twice			
Closed-book test with feedback vs. closed-book test without feedback			
Open-book test vs. closed-book test without feedback			
Open book test vs. Closed book test with feedback			

7. The confidence intervals around the mean difference are somewhat wide. How could the researcher reduce the width of the confidence intervals in future research? Choose two.
   a. Decrease measurement error
   b. Increase the sample size
   c. Increase the treatment effect
   d. Decrease homogeneity of variance

8. What was the overall effect size (η_p^2) for this study? (If you asked SPSS to provide an effect size estimate, you should find this value in the output; if not, you can calculate it by hand easily.)

9. What does the overall effect size tell you?
   a. It indicates how much of an effect studying method variable had on test scores.
   b. It indicates which studying methods were the most helpful.

10. Researchers are usually far more interested in the effect size for the pairwise comparisons than the overall effect size for the ANOVA. Unfortunately, SPSS does not compute the effect sizes for the pairwise comparisons. However, you can compute them by hand using the following formula:

$$d = \frac{M_1 - M_2}{\sqrt{SD_p^2}}.$$

Compute the effect sizes for each pairwise comparison in the following table. The first four pairwise comparisons are completed for you.

	Mean Difference	*Pooled Variance* (SD_p^2)	d
Closed-book test without feedback vs. study twice	$55.33 - 50.83 = 4.5$	$SD_p^2 = \frac{(29)10.25^2 + (29)12.74^2}{(29) + (29)} = 133.63$	.39
Closed-book test with feedback vs. study twice	$65.33 - 50.83 = 14.5$	$SD_p^2 = \frac{(29)11.44^2 + (29)12.74^2}{(29) + (29)} = 146.54$	1.20
Open-book test vs. study twice	$65.67 - 50.83 = 14.84$	$SD_p^2 = \frac{(29)11.43^2 + (29)12.74^2}{(29) + (29)} = 146.43$	1.23
Closed-book test with feedback vs. closed-book test without feedback	$65.33 - 55.33 = 10$	$SD_p^2 = \frac{(29)11.44^2 + (29)10.25^2}{(29) + (29)} = 117.96$	.92
Open-book test vs. closed-book test without feedback			
Open-book test vs. closed-book test with feedback			

11. Write an APA-style summary of the results. When working with this many numbers, it is usually best to put all of the numbers in tables. This has already been done in Tables 1 and 2 below. In your write-up, start with the sentence you completed in Question 1. Then, explain the post hoc tests. In doing so, you should not just write a laundry list of the post hoc tests that are significant and not significant. Instead, you should try to tell a story. Explain what happened in the study. What do the results suggest about how students should study?

Table 1. Means, Standard Deviations, and Confidence Intervals for Each Study Group

Comparison	*Mean* (SD)	*95% CI*
Study twice	50.83 (12.74)	46.08, 55.59
Closed-book test without feedback	55.33 (10.25)	51.51, 59.16
Closed-book test with feedback	65.33 (11.44)	61.06, 69.61
Open-book test	65.67 (11.43)	61.40, 69.93

Table 2. Tukey *HSD* Post Hoc Results With Confidence Intervals and Effect Sizes

Comparison	*Mean Difference*	*95% CI*	p	d
Closed-book test without feedback vs. study twice	4.50	−3.24, 12.24	.431	.39
Closed-book test with feedback vs. study twice	14.50	6.76, 22.24	<.001	1.20
Open-book test vs. study twice	14.83	7.10, 22.57	<.001	1.23
Closed-book test with feedback vs. closed-book test without feedback	10.00	2.26, 17.74	.006	.92
Open-book test vs. closed-book test without feedback	10.33	2.60, 18.07	.004	.95
Open-book test vs. closed-book test with feedback	.33	−7.40, 8.07	.999	.03

Activity 11.6: Choose the Correct Statistic

Learning Objectives

After reading the chapter, completing the homework and this activity, you should be able to do the following:

- Read a research scenario and determine which statistic should be used
- With the addition of one-way independent measures ANOVAs, you now have five different test statistics to choose from. At this point, the one-way ANOVA will be easy to identify because it is the only statistic you know that compares three or more groups. A summary of each possible statistic is provided in the table and the flowchart in the back of the book.
- Like real research, these scenarios can be complex, and some of the information provided is not necessary for deciding which statistic is appropriate. You might find it helpful to circle the information that is important and cross out the information that is not important to make this decision. Try to identify the research goal and then choose the statistic that achieves that goal. Alternatively, try using the flowchart/decision tree at the end of the book to help you focus on the important information in each scenario.

CHOOSE THE CORRECT STATISTIC

Determine which statistic should be used in each of the following research scenarios: z for sample mean, single-sample t, related samples t, independent samples t, or one-way independent samples ANOVA.

1. A psychologist examined the effect of physical exercise on a standardized memory test. Scores on this test for the general population form a normal distribution with a mean of 50 and a standard deviation of 8. A sample of 62 people who exercise at least 3 hours per week has a mean score of 57. Is there evidence for improved memory for those who exercise?

2. A psychologist is interested in the effects of social pressure on conformity behaviors. To investigate the phenomena, she has a subject first sit in a room alone and judge the length of a line. Then, she has the subject sit with confederates who state that the line is much longer than it really is. After the confederates have made their estimates, the subject makes his or hers. The mean length given when they are alone is 5, with a standard deviation of 1.1. The mean length given when they are in a room with confederates is 7, with a standard deviation of 1.9. Which statistic would you use to determine if the length estimates that were provided while the subject was alone were significantly different from those given while the subject was with the confederates?

3. A researcher would like to determine whether the students with anxiety disorders sleep more or less than most students. Suppose it is known that the number of hours college students sleep is normally distributed with a mean of 8. The researcher takes a sample of 32 students with anxiety disorders and records the amount of sleep they get each night. It is found that they average 7.8 hours, with a standard deviation of 1. Which statistic would help determine if students with anxiety disorders sleep more or less than 8 hours?

4. Farmer Brown was curious which of three types of food, Wonder Food, Miracle Grow, or Sorghum, would make cows gain the most weight. He could maximize his profit if he used the feed that was the most effective. But he also had to consider the cost of the food. He would be willing to buy the most expensive feed as long as there was evidence that it was "worth the cost." He purchased a few bags of each type of feed. He randomly selected 30 cows from his herd. He then randomly assigned the cows to be fed Wonder Food, Miracle Grow, or Sorghum. Three months later, he weighed the 30 cows and compared the mean weights of those fed with the different foods. Which statistic would allow you to determine if one of the foods is significantly better than the two?

5. A research scientist for the air force was asked to determine which of two landing gear systems should be installed on a new aircraft that would soon go into production. He had two identical prototypes of the new aircraft and had one of the landing gear systems installed on each aircraft. He obtained the services of 10 highly skilled air force pilots and, by drawing names from a hat, assigned 5 to fly each aircraft. Immediately after each pilot landed, he or she was instructed to rate how well the landing gear performed on a 10-point scale (e.g., 1 = *poorly* and 10 = *very well*). Which statistic would help you determine if one prototype received significantly higher ratings than the other?

6. For a project in their undergraduate stats class, a group of students showed participants a video clip of a car accident. After watching, the video participants had to estimate how fast one of the two cars was going when the accident occurred. A third of the participants were asked, "How fast was the car going when it *bumped* the other car?" A third were asked, "How fast was the car going when it *hit* the other car?" And the final third were asked, "How fast was the car going when it *smashed* the other car?" Which statistic would enable you to determine whether the three different ways of asking the question lead to different estimates of the car's speed?

7. Executives at a large corporation hired a psychologist to evaluate two physical fitness programs with the intent of adopting one for their employees. The two fitness programs were creatively labeled Plan A and Plan B. Two different work teams (employees who work together) volunteered to participate in some kind of fitness program. These two work teams were randomly assigned to one of the two different fitness plans. After 6 months, the psychologist used corporation records to determine the rate of absenteeism (number of work days missed) for employees in each plan. Which statistic would help you determine if Plan A or Plan B leads to less absenteeism?

8. A researcher studies the effect of a drug on the number of nightmares occurring in veterans with posttraumatic stress disorder (PTSD). A sample of clients with PTSD kept records of each incident of a nightmare for 1 month before treatment. Subjects were then given the medication for 1 month, and they continued to record each nightmare. Which statistic would help you determine if the medication significantly reduced nightmares?

9. A college professor has noted that this year's freshman class appears to be smarter than classes from previous years. The professor obtains a sample of $N = 36$ freshmen and computes an average IQ ($M = 114.5$) and the variance in IQ ($SD^2 = 324$) for the sample. College records indicate that the mean IQ for the population of entering freshman from previous years was 110.3. Is the incoming class significantly smarter (as measured by IQ) than the mean for previous years?

10. A group of participants had to press the brake on a driving simulator whenever they saw a ball roll into the road ahead of them. Every driver performed the task while talking on a cell phone and while not talking on a cell phone. The DV was how long in milliseconds it took to press the

brake pedal. The researchers compared the mean reaction time in the cell phone condition with the mean reaction time in the control condition. Which statistic would help determine if talking on a cell phone leads to slower response times?

11. A neurologist had two groups of patients with different kinds of aphasia (i.e., brain disorders). One group had a damage to Broca's area and the other had damage to Wernicke's area. The researcher also used a third group of people who had no brain abnormalities. People in each group were shown line drawings of common household objects (i.e., a lamp, a chair, etc.) and were asked to name the object. Which statistic would determine if the number of objects correctly identified differed across the three groups?

Chapter 11 Practice Test

A psychologist would like to determine which of three therapies is most effective in treating agoraphobia: (1) CBT, (2) systematic desensitization, or (3) hypnosis. Fifteen individuals diagnosed with agoraphobia were randomly assigned to each of the three therapies. After 3 months of therapy, the psychologist asks the participants to record the number of times they leave their house/apartment building during a 2-week period.

#1	#2	#3
Cognitive Behavioral Therapy	*Systematic Desensitization*	*Hypnosis*
7	12	2
8	7	3
6	8	1
9	9	4
10	6	2

1. Match the assumption to the fact that is relevant to that assumption.

   C Independence
   A Appropriate measurement of the IV and the DV
   B Normality
   D Homogeneity of variance

   a. The one variable defines groups and the participants' responses were given on an interval/ratio scale.
   b. Samples of 30 or more tend to form distributions of sample means that meet this assumption; also, if the population of scores has a normal shape, this assumption will be met.
   c. Data were collected from one participant at a time.
   d. When the standard deviations from each condition are similar (i.e., not twice as large), this assumption is probably met.

2. Which of the following best represents the null hypothesis for this study?

   a. H_0: $\mu_1 = \mu_2 \neq \mu_3$
   b. H_0: $\mu_1 \neq \mu_2 = \mu_3$
   c. H_0: $\mu_1 \neq \mu_2 \neq \mu_3$
   d. H_0: $\mu_1 = \mu_2 = \mu_3$

3. Which of the following best represents the *research hypothesis* for this study?

   a. The mean number of times patients with agoraphobia leave their homes after treatment is the same for CBT, systematic desensitization, and hypnosis.
   b. The mean number of times patients with agoraphobia leave their homes after treatment is not the same for CBT, systematic desensitization, and hypnosis.

The following ANOVA source table came from the preceding study, but it is only partly complete. Complete the entire table and use it to answer the next three questions. (Complete the table by hand *before* you analyze the data using SPSS.)

Source of Variance	SS	df	MS	F
Between	112.53	2	56.265	18.548
Within (error)	36.40	12	3.033	
Total	148.93			

4. What is the between-treatments *df* for this study with three groups (i.e., three conditions)?
   a. 2
   b. 3
   c. 4

$df_b = g - 1$
$3 - 1$

5. What is the within (error) *df* for this study with five people in each of the three groups?
   a. 4
   b. 5
   c. 14
   d. 12

$df_W = N - g$
$= 15 - 3$
$= 12$

6. What is the error term (i.e., $MS_{within(error)}$)?
   a. 36.40
   b. 3.033
   c. 56.267
   d. 18.549

$MS = \frac{SS_W}{df_W} \rightarrow \frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{N - g}$

7. What is the critical value of *F* for this study that is using an alpha value of .05?
   a. 3.89
   b. 3.68
   c. 3.49

$D = 12$
$N = 2$

8. What is the obtained *F* value for this study?
   a. 36.40
   b. 3.033
   c. 56.267
   d. 18.549
   e. 194.198

9. Should the null hypothesis of this study be rejected?
   a. Yes
   b. No

10. Do post hoc tests need to be done for this study?
    a. Yes, we failed to reject the null and there are only three groups.
    b. Yes, we rejected the null and there are more than two groups.
    c. No, we rejected the null and there are more than two groups.
    d. No, we failed to reject the null and there are only three groups.

11. Which of the following images shows the correct way to enter the independent ANOVA data into SPSS for this analysis?

Data Set A

Cognitive	Systematic	Hyponosis
7.00	12.00	2.00
8.00	7.00	3.00
6.00	8.00	1.00
9.00	9.00	4.00
10.00	6.00	2.00

Data Set B

Treatment	LeaveHouse
1.00	7.00
1.00	8.00
1.00	6.00
1.00	9.00
1.00	10.00
2.00	12.00
2.00	7.00
2.00	8.00
2.00	9.00
2.00	6.00
3.00	2.00
3.00	3.00
3.00	1.00
3.00	4.00
3.00	2.00

The SPSS output for the ANOVA and post hoc tests follows. Use this output to answer questions.

Descriptive Statistics

Dependent Variable: LeftHouse

Treatment	Mean	Std. Deviation	N
cbt	8.0000	1.58114	5
sd	8.4000	2.30217	5
hyp	2.4000	1.14018	5
Total	6.2667	3.26161	15

Tests of Between-Subjects Effects

Dependent Variable: LeftHouse

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	112.533[a]	2	56.267	18.549	.000	.756
Intercept	589.067	1	589.067	194.198	.000	.942
Treatment	112.533	2	56.267	18.549	.000	.756
Error	36.400	12	3.033			
Total	738.000	15				
Corrected Total	148.933	14				

a. R Squared = .756 (Adjusted R Squared = .715)

Multiple Comparisons

Dependent Variable: LeftHouse

Tukey HSD

(I) Treatment	(J) Treatment	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
cbt	sd	-.4000	1.10151	.930	-3.3387	2.5387
	hyp	5.6000*	1.10151	.001	2.6613	8.5387
sd	cbt	.4000	1.10151	.930	-2.5387	3.3387
	hyp	6.0000*	1.10151	.000	3.0613	8.9387
hyp	cbt	-5.6000*	1.10151	.001	-8.5387	-2.6613
	sd	-6.0000*	1.10151	.000	-8.9387	-3.0613

Based on observed means.
The error term is Mean Square(Error) = 3.033

*. The mean difference is significant at the .05 level.

12. What is η_p^2 for the effect of the treatments on the dependent variable?
    a. .756
    b. .942
    c. 18.549
    d. 194.198

13. Which of the post hoc tests are significant? Select all that are significant.
    a. CBT vs. SD
    b. CBT vs. HYP
    c. SD vs. HYP

14. Which of the following accurately describes the results of the Tukey post hoc tests?
    a. Systematic desensitization was significantly better than cognitive behavioral therapy and hypnosis. Cognitive behavioral therapy and hypnosis were the same.
    b. Systematic desensitization and cognitive behavioral therapy were equally effective. Hypnosis was worse than both of them.
    c. Hypnosis was the best, cognitive behavioral therapy was the second best, and systematic desensitization was the worst.

15. The effect sizes for two of the pairwise comparisons have been computed and entered in the following table. Compute the remaining effect size.

Table 1 Means and Standard Deviations for Treatment

Class	*Mean (SD)*
Cognitive behavioral	8.00 (1.58)
Systematic desensitization	8.40 (2.30)
Hypnosis	2.40 (1.14)

Table 2 Tukey *HSD* Post Hoc Results With Effect Sizes

Comparison	*Mean Difference*	p	d
CBT vs. SD	−.40		
CBT vs. HYP	5.60	.001	4.06
SD vs. HYP	6.00	<.001	3.30

16. What is the p value for the CBT vs. SD post hoc comparison?
    a. .001
    b. .930
    c. .05
    d. .869

17. Why do you need to compute two measures of effect size (d and η_p^2) for ANOVAs?
    a. η_p^2 allows you to quantify the size of the overall effect of the IV on the DV. d allows you to quantify the size of the differences between each pair of means.
    b. η_p^2 allows you determine which IV is more effective. d allows you to determine which DV is more effective.
    c. η_p^2 allows you determine whether you should do post hoc tests. d tells you which post hoc tests are significantly different than 0.

18. Which of the following is the best summary of the results of this study?
    a. A one-way ANOVA was conducted with type of treatment as the independent variable and the number of times each person left the house as the DV. The effect of treatment was significant, $F(2, 12) = 18.55$, $p < .001$, $MSE = 3.03$, $\eta_p^2 = .76$. Tukey *HSD* post hoc tests indicated that participants who received cognitive behavioral therapy left the house more often than participants who received hypnosis. No other differences were significant. See Tables 1 and 2 for descriptive statistics.
    b. A one-way ANOVA was conducted with type of treatment as the independent variable and the number of times each person left the house as the DV. The effect of treatment was significant, $F(2, 12) = 18.55$, $p < .001$, $MSE = 3.03$, $\eta_p^2 = .76$. Tukey *HSD* post hoc tests indicated that participants who received cognitive behavioral therapy left the house more often than participants who received hypnosis. Likewise, participants who received systematic desensitization left the house more often than participants who received hypnosis. Cognitive behavioral therapy and systematic desensitization were equally effective. See Tables 1 and 2 for descriptive statistics.
    c. A one-way ANOVA was conducted with type of treatment as the independent variable and the number of times each person left the house as the DV. The effect of treatment was significant, $F(2, 12) = 18.55$, $p < .001$, $MSE = 3.03$, $\eta_p^2 = .76$. Tukey *HSD* post hoc tests indicated that participants who received systematic desensitization left the house more often than participants who received cognitive behavioral therapy who left the house more often than participants who received hypnosis. Overall, systematic desensitization was the most effective treatment. See Tables 1 and 2 for descriptive statistics.
    d. A one-way ANOVA was conducted with type of treatment as the independent variable and the number of times each person left the house as the DV. The effect of treatment was not significant, $F(2, 12) = 18.55$, $p < .001$, $MSE = 3.03$, $\eta_p^2 = .76$. Overall, all treatments were equally effective. See Tables 1 and 2 for descriptive statistics.

Use the following two data sets to answer Questions 19 through 23.

Researcher 1		
Group 1: Cola	*Group 2: Club Soda*	*Group 3: Water*
48	56	64
50	60	70
52	64	76
$M_1 = 50$	$M_2 = 60$	$M_3 = 70$
$SD_1 = 2$	$SD_2 = 4$	$SD_3 = 6$
$SS_1 = 8$	$SS_2 = 32$	$SS_3 = 72$

Researcher 2		
Group 1: Cola	*Group 2: Club Soda*	*Group 3: Water*
40	40	40
50	60	70
60	80	100
$M_1 = 50$	$M_2 = 60$	$M_3 = 70$
$SD_1 = 10$	$SD_2 = 20$	$SD_3 = 30$
$SS_1 = 200$	$SS_2 = 800$	$SS_3 = 1{,}800$

19. Which researcher's data set has more variability within treatments (i.e., a higher $MS_{within(error)}$)?
    a. Researcher 1
    b. Researcher 2
    c. They have the same amount of within-treatments variability.
20. Which researcher's data set has more variability between treatments (i.e., a higher $MS_{between}$)?
    a. Researcher 1
    b. Researcher 2
    c. They have the same amount of between-treatments variability.
21. Which researcher's data sets will result in a larger F value?
    a. Researcher 1
    b. Researcher 2
    c. They have the same F value.
22. Which researcher's study is more likely to lead to rejecting the null hypothesis?
    a. Researcher 1
    b. Researcher 2
23. Which data set probably has more measurement error?
    a. The data set of Researcher 1
    b. The data set of Researcher 2

A researcher reported the following results: $F(2, 72) = 4.00$, $p < .05$, $MSE = 6$. Use this information to "work backward" and complete the following source table.

Source	SS	df	MS	F	η_p^2
Between treatments					
Within treatments					
Total					

24. How many treatment conditions (groups) were in this study?
    a. 1
    b. 2
    c. 3
    d. 4

25. How many people participated in this study?
    a. 73
    b. 74
    c. 75
    d. 76

26. What is the critical value for the *F* (using an alpha of .05)?
    a. 4.00
    b. 3.13
    c. 2.73
    d. 1

27. Should you reject or fail to reject the null hypothesis?
    a. Reject
    b. Fail to reject

28. Do post hoc tests need to be done?
    a. Yes
    b. No

REFERENCES

Agarwal, P. K., Karpicke, J. D., Kang, S. K., Roediger, H. I., & McDermott, K. B. (2007). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22*(7), 861–876.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Field, A. P. (2013). *Discovering statistics using IBM SPSS Statistics: And sex and drugs and rock 'n' roll* (4th ed.). London, England: Sage.

CHAPTER 12

Two-Factor ANOVA or Two-Way ANOVA

LEARNING OBJECTIVES

After reading this chapter, you should be able to do the following:

- Explain when to use a two-factor analysis of variance (ANOVA)
- Describe the three *F* ratios generated by a two-factor ANOVA
- Write null and research hypotheses for the main effects and interaction *F* tests
- Identify which means are compared when computing each main effect and interaction *F* tests
- Complete an ANOVA summary table
- Define the critical region for each *F* test
- Determine whether you should reject each null hypothesis
- Compute effect sizes for each main effect and the interaction, and describe each as small, medium, or large
- Use SPSS to perform a two-factor ANOVA and create a graph of the main effects and interaction
- Use SPSS to compute simple effect analyses
- Summarize the results of the ANOVA using American Psychological Association (APA) style
- Interpret the SPSS output for a two-factor ANOVA

PURPOSE OF TWO-WAY ANOVA

In the previous chapter, you learned about the one-way, independent ANOVA (also called the single-factor independent measures ANOVA). This statistic determines if a single independent variable (IV) has a significant effect on a dependent variable (DV). When conducting ANOVAs, IVs are often referred to as factors; therefore, an ANOVA evaluating a single IV is called a single-factor ANOVA or a one-way ANOVA. This chapter introduces the two-factor ANOVA, or the two-way ANOVA. Two-way ANOVAs are used frequently in the behavioral sciences because they can test for the effects of two IVs/two factors

at the same time, and more important, two-way ANOVAs allow you to determine if the two IVs *interact* to affect the DV. The ability to examine the joint effect of two IVs (i.e., their interaction effect on a DV) is what makes the two-way ANOVA such a useful research tool. The world is complex, and multiple factors are important in most situations. The two-way ANOVA enables you to design studies that more closely mimic the complexity of the real world.

Reading Question

1. A two-way ANOVA (also called a two-factor ANOVA) has two
   a. IVs.
   b. DVs.

Reading Question

2. A two-way ANOVA (also called a two-factor ANOVA) allows researchers to determine if
   a. two IVs interact to jointly affect a DV.
   b. an IV has a large, medium, or small effect on a DV.

DESCRIBING FACTORIAL DESIGNS

For example, suppose you wanted to know the best way to study for college exams. Should you read over your notes or should you "test yourself" by trying to recall information from your notes? In your cognitive psychology course, you heard about the "testing effect," which suggests that the act of trying to recall information aids memory. But, you have studied for a lot of tests in your academic career; you have always just read over your notes, and it seems to work. Might self-testing be a better way to study? You decide to conduct an experiment for your senior project in which you manipulate the way participants study; half will study by *rereading* material, and half will study by *trying to recall* material they read previously. Thus, type of studying method is an IV in your study, and it has two **levels**: either *rereading* material or *trying to recall* material. You also remember from your cognition class that human memory performance is highly influenced by time. In other words, the length of the retention period is an additional important factor to consider. It is possible that the best way to study for an exam might depend on how much time there is until the exam. So, you decide to include a second IV in your study. Specifically, you will manipulate the time delay between the study session and the exam; half of the participants will take an exam 5 minutes after studying, and half will take an exam 2 days later. Thus, time delay is a second IV in your study, and it has two levels: either *5-minute delay* or *2-day delay*. Table 12.1 will help you visualize the design of your study.

Table 12.1 Design of Two-Factor Study

		Factor A—Studying Method	
		Rereading	*Trying to Recall*
Factor B—Time Delay	5 minutes		
	2 days		

The two IVs in your study combine to create four experimental conditions. One group of participants will reread their notes *and* take the exam 5 minutes later. A second group will reread their notes *and* take the exam 2 days later. A third group will try recalling material *and* take the exam 5 minutes later. And the final group will try recalling material *and* take the exam 2 days later. The design you created is a 2 × 2 between-group factorial design. It is a 2 × 2 because the first IV, studying method, has two levels (rereading or recalling), and the second IV, time delay, has two levels (5 minutes or 2 days). If you had decided to have three levels of the variable time delay (e.g., 5 minutes, 2 days, or 1 week), your design would have been a 2 × 3 between-group factorial design. This larger design would have had six conditions (2 × 3 = 6), sometimes called "cells," rather than four conditions, or "cells." So, whenever you describe a two-way factorial design, you will need two numbers, each reflecting the number of levels one of the IVs has in that design. Both of these designs are between-group designs because there are different participants in each of the different conditions, or cells. These types of designs are also referred to as independent samples or independent measures designs.

Reading Question

3. Participants take a test 5 minutes or 2 days after studying. How many *variables* does this information represent?
   a. 1
   b. 2

Reading Question

4. Participants take a test 5 minutes or 2 days after studying. How many *levels* does this variable have?
   a. 1
   b. 2

Reading Question

5. How many cells are in a 3 × 3 between-group ANOVA?
   a. 2
   b. 3
   c. 6
   d. 9

Reading Question

6. How many independent variables are in a 3 × 3 between-group ANOVA?
   a. 2
   b. 3
   c. 6
   d. 9

LOGIC OF THE TWO-WAY ANOVA

Two-way ANOVAs actually produce *three different significance tests,* each with its own null hypothesis. Your understanding of how the two-way ANOVA produces these three tests will hinge on your understanding of Table 12.2. This table displays a more detailed depiction of the 2 × 2 between-group design you created previously.

Table 12.2 2 × 2 ANOVA With Cell Means and Marginal Means

		Factor A—Studying Method		
		Rereading	*Trying to Recall*	
Factor B—Time Delay	5 minutes	8, 6, 11, 7, 5, 9, 10, 8 Cell mean = 8.00 (2.00)	7, 5, 10, 6, 5, 9, 9, 7 Cell mean = 7.25 (1.91)	Marginal mean for 5 minutes = 7.63 (1.93)
	2 days	5, 3, 7, 5, 2, 6, 6, 6 Cell mean = 5.00 (1.69)	7, 5, 8, 7, 4, 8, 9, 8 Cell mean = 7.00 (1.69)	Marginal mean for 2 days = 6.00 (1.93)
		Marginal mean for rereading = 6.50 (2.37)	Marginal mean for trying to recall = 7.13 (1.75)	

In your 2 × 2 between-group design, you randomly assign each of your 32 research participants to one of the four different conditions, or cells. Consequently, you have eight participants in each of the four cells. All of these participants read a passage from a college textbook and then studied for a future exam by either rereading the passage or trying to recall the information they read. Then, half of those who reread took a subsequent exam 5 minutes later and half took the same exam 2 days later. Similarly, half of those who studied by trying to recall information they read took the exam 5 minutes after studying, and the other half took the exam 2 days later. Each participant's number of correct answers on the exam is listed in each of the above cells. In addition, the mean for each cell is also provided in Table 12.2; these values are called **cell means,** but it is also correct to call them **condition means.**

As stated previously, the two-way ANOVA creates three significance tests. One of these tests is the **main effect of Factor A (study method).** In the study you designed, this test compares the mean DV scores of *everyone* who studied by rereading to the mean of *everyone* who studied by recalling. In other words, the main effect of studying method compares the marginal means of rereading ($M = 6.50$) and recalling ($M = 7.13$) from Table 12.2. The logic of this test is identical to that of a one-way ANOVA. A *conceptual* formula of this analysis is provided as follows:

$$F_{\text{Studying Method}} = \frac{\text{Studying Method effect \& individual differences \& measurement error}}{\text{Individual differences \& measurement error}}.$$

The numerator of this test is the average between-group variability, which represents variability created by some combination of the treatment effect (i.e., studying method), individual differences, and measurement error. The denominator of this test is the average within-group variability, which represents the variability created by individual differences and measurement error (i.e., sampling error). If the different studying methods are equally effective, if they create no variability, then the obtained F value for this main effect would equal 1. Conversely, if one studying method is more

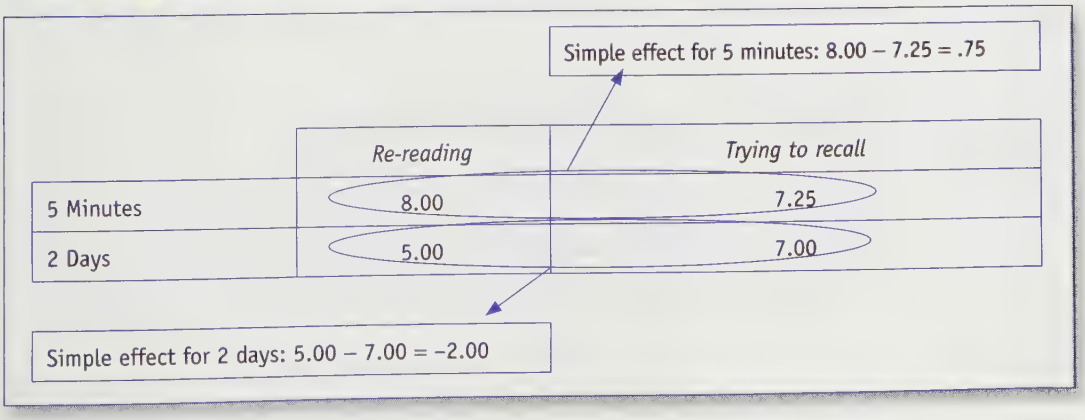
effective than the other, if there is variability created by the different methods, the obtained F value will be substantially larger than 1. Hopefully, you recognize that this logic is identical to that of a one-way ANOVA. In fact, the obtained F value of this main effect would be identical to the obtained F value of a one-way ANOVA.

This two-way ANOVA also creates a test for the main effect of Factor B (time delay). In the study you designed, this test compares the mean DV score of *everyone* who experienced a 5-minute delay to the mean DV score of *everyone* who experienced a 2-day delay. This main effect compares the marginal means of a 5-minute delay ($M = 7.63$) and a 2-day delay ($M = 6.00$) in Table 12.2. Again the logic is identical to a one-way ANOVA. This test is different from the main effect of Factor A because the data are grouped differently (which explains why the marginal means compared by this test are different from those compared by the main effect of study method). The conceptual formula for this main effect of time delay is provided below. Again, if the different time delays create no variance, the obtained F value should be equal to 1. As before, if the different time delays do create variance, the obtained F value will be substantially larger than 1.

$$F_{\text{Time Delay}} = \frac{\text{Time Delay effect \& individual differences \& measurement error}}{\text{Individual differences \& measurement error}}.$$

Finally, this two-way ANOVA also creates a test for the interaction effect of Factor A and Factor B (Study Method × Time Delay). While the main effects compare the differences between their respective marginal means, the interaction compares the differences between simple effects. Simple effects *are the differences between pairs of cell means*. For example, in your study, when participants experienced a 5-minute delay, the difference between the cell means was 8.00 – 7.25 = 0.75. This difference, 0.75, is a simple effect. It indicates that, when the time delay was 5 minutes, those who reread recalled .75 *more* correct answers than those who tried recalling. When participants experienced a 2-day delay, the simple effect was 5.0 −7.0 = −2.0. It indicates that, when the time delay was 2 days, those who reread recalled two *less* correct answers than those who tried recalling. Table 12.3 may help you visualize these simple effects.

Table 12.3 Simple Effects for 2 × 2 ANOVA



The interaction effect compares these two simple effects (0.75 vs. −2.0); the greater the difference between these simple effects, the greater the *F* value will be for this interaction. The conceptual formula for this interaction effect is as follows:

$$F_{\text{Method Delay}} = \frac{\text{Interaction Studying Method} \times \text{Time Delay effect \& ind. diff. \& measurement error}}{\text{Individual differences \& measurement error}}.$$

This interaction effect is typically the most important of the three tests provided by the two-way ANOVA, so it is essential that you understand what it is testing. Generally speaking, the interaction effect is testing if the two IVs combine to influence the DV. In the study you designed, the interaction helps you determine if students should study differently for exams depending on how far away the exam is. *If* there is a significant interaction, the answer to "How should I study?" *depends* on how much time there is before the exam. If you look at the simple effects, you can get a preview of what the interaction *might* tell you. The simple effect for the 5-minute delay condition (i.e., 8.0 − 7.25 = 0.75) suggests that the rereading participants (8.0) did slightly better than the recalling participants (7.5); however, the simple effect for the 2-day delay condition (i.e., 5.0 − 7.0 = −2.0) suggests something completely different—namely, that the recalling participants (7.0) did quite a bit better than the rereading participants (5.0). The fact that the two simple effects suggest different "stories" or different answers to "How should I study?" is an indication that there *might* be a meaningful (i.e., statistically significant) interaction between studying method and time delay. The obtained *F* value for the interaction effect will indicate whether the difference between the simple effects (i.e., 0.75 vs. −2.0) is likely to have resulted from sampling error. If the obtained *F* value is close to 1 (defined by the critical value of *F*), then the simple effect differences are considered to be due to sampling error.

The remainder of the chapter describes the details you would need to analyze the 2 × 2 between-group factorial design you created to answer the question "How should I study for college exams?" Your study is very similar to one conducted by Roediger and Karpicke (2006). The pattern of data and conclusions described below are consistent with those of Roediger and Karpicke (2006).

Reading Question

7. How many main effects are tested by a two-factor ANOVA?
   a. 1
   b. 2
   c. 3

Reading Question

8. How many interactions are tested by a two-factor ANOVA?
   a. 1
   b. 2
   c. 3

Reading Question

9. How many *F* ratios are computed with a two-factor ANOVA?
   a. 1
   b. 2
   c. 3

Reading Question

10. A test of a main effect compares the _____ to determine if they are significantly different from each other.
    a. simple effects
    b. cell means
    c. marginal means

Reading Question

11. A test of an interaction compares the _____ to determine if they are significantly different from each other.
    a. simple effects
    b. cell means
    c. marginal means

Reading Question

12. Which analysis helps you determine if the effect of one IV depends on a different IV?
    a. Main effect
    b. Interaction

EXAMPLE OF A TWO-WAY ANOVA

Step 1: Examine Variables to Assess Statistical Assumptions

The statistical assumptions for the two-way ANOVA are similar to those for the one-way ANOVA. There are four assumptions to examine. The responses within each condition must not be influenced by other responses within that same condition (*data independence*), the distribution of sample means for each condition (i.e., cell) must have a normal shape (*normality*), the variability within each cell should be similar (*homogeneity of variance*), and the variables being analyzed must be appropriate for the two-way ANOVA (*appropriate measurement of variables*). The procedural controls used in this study seem likely to provide data independence. As with previous examples, the sample sizes in this example are too small for a real research study, but if we assume the populations represented by each condition are normally distributed, this example will be sufficient for teaching purposes. The homogeneity of variance assumption is satisfied because none of the conditions' standard deviations are double the size of any other conditions. The two-way ANOVA requires two "grouping" IVs and a DV that is measured on an interval or ratio scale. The first IV in this study, studying method, identifies two groups of participants: One studies by rereading and the other by trying to recall. The second IV, time delay, identifies two groups that test 5 minutes after studying or 2 days after studying. Together, these two variables create four conditions: rereading/5-minute delay, rereading/2-day delay, recalling/5-minute delay, and recalling/2-day delay. Whenever the conditions defined by two IVs are "combined" to create all possible combinations of IV conditions, it is a good indication that the two-way ANOVA is the appropriate statistic for those data. The DV in your study was the number of correct answers on an exam, which is an interval/ratio variable. So, a two-way ANOVA is the correct statistic for analyzing your data.

Reading Question

13. You use a two-way ANOVA when
    a. the IV defines two independent samples and the DV is measured on an interval/ratio scale.

b. the IV defines two or more independent samples and the DV is measured on an interval/ratio scale.

c. the IV defines two matched samples and the DV is measured on an interval/ratio scale.

d. there are two IVs that combine to define four or more independent samples and the DV is measured on an interval/ratio scale.

e. the IV defines one sample, the DV is measured on an interval/ratio scale, and the DV is measured twice on that same sample.

f. the IV defines one sample and the DV is measured on an interval/ratio scale, and you do not know the population standard deviation.

g. the IV defines one sample and the DV is measured on an interval/ratio scale, and you do know the population standard deviation.

Step 2: Set Up the Null and Research Hypotheses

This two-factor ANOVA tests for three different effects. Specifically, it determines (1) if studying method has an effect on exam scores, (2) if time delay until the exam has an effect on exam scores, and (3) if studying method and time delay interact to jointly affect exam scores. If the two variables interact, it means that the best studying method to use depends on the time you have until the exam. Each of these three effects can be computed or discussed in any order, but we prefer to deal with the interaction effect first because it is generally the most important of the three effects. If you end up with a meaningful (i.e., statistically significant) interaction, you will probably want to tell people about it first because it is the most important finding. If, however, you don't have a significant interaction, you might want to talk about a significant main effect first when you report your results. We will talk about how to report the results later in the chapter, but in each of the following steps, we will discuss the test for the interaction effect first because generally that is what you are most interested in when you conduct a two-way ANOVA.

The two-way ANOVA generates three different effects, so there are three different sets of null and research hypotheses. We will start with the hypotheses for the interaction and then describe those for the two main effects.

2a. Interaction Null and Research Hypotheses

The primary advantage in using a two-factor design is that it allows you to determine if studying method and time delay until an exam interact to affect exam scores. An interaction between these IVs exists if the effect studying method has on exam scores is different when the time until the exam is relatively short (e.g., 5 minutes) versus relatively long (e.g., 2 days). Thus, the interaction research hypothesis is that the effect of studying method is different for people with short time delays versus long time delays. The interaction null hypothesis is that the effect of studying method is the same no matter when the exam takes place. The interaction hypotheses are summarized in Table 12.4.

Reading Question

14. The null hypothesis for the interaction states that the effect of studying method on exam scores

a. is the same for people with short and long periods between studying and the exam.

b. is different for people with short and long periods between studying and the exam.

Table 12.4 Verbal Representations of the Research and Null Hypotheses for the Study Method × Time Delay Interaction

	Verbal Statement of Hypothesis	*Mean Difference in Simple Effects Created by*
Research hypothesis (H_1)	The effect of studying method on exam scores is different for people with short versus long time delays between studying and the exam (i.e., there is an interaction between the two IVs).	The IVs interacting to affect memory performance
Null hypothesis (H_0)	The effect of studying method on exam scores is the same for people with short and long time delays between studying and the exam (i.e., there is no interaction between the two IVs).	Sampling error

Reading Question

15. The research hypothesis for the interaction states that the effect of studying method on exam scores
    a. is the same for people with short and long periods between studying and the exam.
    b. is different for people with short and long periods between studying and the exam.

As described above, there might be an interaction if the simple effects created by the cell means are very different. In the previous section, we generated the set of simple effects for time delay by finding the difference between the cell means for a 5-minute delay (8.00 – 7.25 = 0.75) and the difference for a 2-day delay (5.00 – 7.00 = –2.00). These generated simple effects were different by 2.75; +0.75 is 2.75 away from –2.00. What if we had decided to create the simple effects for studying method? Could using a different set of simple effects potentially change whether or not we find an interaction? The simple effect for rereading would be 3.00 (8.00 – 5.00 = 3.00), and the simple effect for recalling would be 0.25 (7.25 – 7.0 = 0.25). These simple effects are also different by 2.75 (3.00 – 0.25 = 2.75). So, it does not matter which set of simple effects one chooses to use; both will produce the same conclusion concerning the interaction. You should choose the set of simple effects that make the most sense in terms of your theory or the research literature. Essentially, you can choose the set of simple effects that makes the most sense to you.

The simple effects for time delay indicate that when there was a 5-minute time delay, participants who studied by rereading (M = 8.00) had slightly more correct than those who studied by recalling (M = 7.25). However, when the time delay was 2 days, the rereaders had fewer correct (M = 5.00) than the recallers (M = 7.00). Now, you need a significance test to determine if this pattern of cell means (i.e., the difference between the simple effects) is likely to have occurred due to sampling error. As with all previous significance tests, not rejecting the null hypothesis suggests that the differences were created by sampling error.

Reading Question

16. The null hypothesis for the interaction compares pairs of _____ that are generated by the difference between the _____.
    a. main effects; marginal means

b. simple effects; cell means
c. marginal effects; condition means

Reading Question

17. To determine if an interaction is *statistically significant,* you will have to
a. interpret the cell means.
b. compute a significance test.

2b. Main Effect Null and Research Hypotheses

The second effect is the main effect of studying method. When you are analyzing this effect, you are just analyzing the IV of studying method. You are essentially ignoring the second IV of time delay. The **null hypothesis** for the main effect of studying method is that participants who studied by rereading versus trying to recall information from the reading will have equivalent exam scores. The **research hypothesis** is that these two conditions will have different exam scores. The symbolic notations for these hypotheses as well as their verbal equivalents are provided in Table 12.5.

Reading Question

18. When you are investigating the effect of one IV on the DV, the effect is called
a. a main effect.
b. an interaction effect.

Reading Question

19. The null hypothesis for the main effect of studying method states that
a. the different studying methods will produce similar exam scores.
b. the different studying methods will produce different exam scores.

Reading Question

20. The research hypothesis for the main effect of studying method states that
a. the different studying methods will produce similar exam scores.
b. the different studying methods will produce different exam scores.

Table 12.5 Symbolic and Verbal Representations of the Research and Null Hypotheses for the Main Effect of Study Method

	Symbolic	*Verbal*	*Mean Difference Created by*
Research hypothesis (H_1)	H_1: $\mu_1 \neq \mu_2$	The exam scores for those using the different studying methods will not be equal.	One studying method being better than the other
Null hypothesis (H_0)	H_0: $\mu_1 = \mu_2$	The exam scores for those using the different studying methods will be equal.	Sampling error

You should understand that this main effect is comparing the mean for *all participants* who studied by rereading to the mean for *all participants* who studied by trying to recall. These two means are the marginal means for studying method. When there are the same number of participants in each cell of the design, the rereading marginal mean is the mean of all participants who reread, specifically, the mean of the rereading/5-minute (M = 8.00) and rereading/2-day (M = 5.00) cell means. The mean of 8.00 and 5.00 is 6.50, so the mean exam score for all participants in the rereading condition is 6.50 correct answers. This average of the two rereading cell means is typically written on the margin of the table, so it is referred to as a marginal mean (see Table 12.6).

Table 12.6 Marginal Means for the Main Effect of Studying Method

	Studying Method	
	Rereading	*Trying to Recall*
5-minute delay	8.00	7.25
2-day delay	5.00	7.00
Marginal means	**6.50**	**7.13**

Similarly, the mean exam score for all participants who studied by trying to recall material is the mean of the 5-minute/trying to recall (M = 7.25) and 2-day/trying to recall (M = 7.00) cell means. In this case, the marginal mean is 7.13, and so the mean exam score for all those who studied by recalling material is 7.13.

Later, you will compute an F value to determine if the main effect of studying method is statistically significant. In other words, you will determine if the difference between the marginal means of 6.50 correct answers and 7.13 correct answers is likely to have been created by sampling error or not. If the difference is not likely to be created by sampling error, the mean difference was probably created by differences in studying methods.

Reading Question

21. To test for the main effect of studying method, you will compute an _____ to determine if the marginal means of _____ and _____ are significantly different.
    a. F test; 8.00; 7.25
    b. F test; 6.50; 7.13

2c. The Other Main Effect Null and Research Hypotheses

The last test is the main effect of time delay until the exam. The null hypothesis for this main effect is that participants who experienced a 5-minute delay versus a 2-day delay will generate the same number of correct answers on the exam. The research hypothesis is that these two groups will generate different numbers of correct answers. The symbolic and verbal representations of the hypotheses are given in Table 12.7.

Reading Question

22. The null hypothesis for the main effect of time delay predicts that
    a. the mean number of correct answers for the 5-minute delay group will be equal to the mean number of correct answers for the 2-day delay group.
    b. the mean number of correct answers for the 5-minute group will be different from the mean number of correct answers for the 2-day delay group.

Table 12.7 Symbolic and Verbal Representations of the Research and Null Hypotheses for the Main Effect of Time Delay

	Symbolic	*Verbal*	*Mean Difference Created by*
Research hypothesis (H_1)	H_1: $\mu_1 \neq \mu_2$	The mean of the population of exam scores produced after a 5-minute delay *is not equal* to the mean of the population of exam scores produced after a 2-day delay.	One time delay affecting memory more than the other
Null hypothesis (H_0)	H_0: $\mu_1 = \mu_2$	The mean of the population of exam scores produced after a 5-minute delay *is equal* to the mean of the population of exam scores produced after a 2-day delay.	Sampling error

Reading Question

23. The research hypothesis for the main effect of time delay predicts that
    a. the mean number of correct answers for the 5-minute delay group will be equal to the mean number of correct answers for the 2-day delay group.
    b. the mean number of correct answers for the 5-minute group will be different from the mean number of correct answers for the 2-day delay group.

Table 12.8 Marginal Means for the Main Effect of Time Delay

	Rereading	*Trying to Recall*	*Marginal Means*
5-minute delay	8.00	7.25	**7.63**
2-day delay	5.00	7.00	**6.00**

This main effect is comparing the mean exam score of all participants who had a 5-minute delay to the mean of all participants who had a 2-day delay. As can be seen in Table 12.8, the marginal mean for all of the 5-minute delay participants is 7.63, the mean of 8.00 and 7.25, which were the cell means for the two groups of 5-minute delay participants. Similarly, the marginal mean for the 2-day delay participants is 6.00, the mean of 5.00 and 7.00, which were the cell means for the two groups of 2-day delay participants. Later, you will compute an *F* value to determine if the difference between 7.63 correct answers and 6.00 correct answers is statistically significant (i.e., if the difference is unlikely to be caused by sampling error).

Reading Question

24. Which two means are being compared when you compute the *F* test for the main effect of time delay?
    a. 7.25 and 7.00
    b. 7.63 and 6.00

Step 3: Define the Critical Region

As you learned in the previous section, the two-way ANOVA produces three *F* tests. Each test has its own critical region. To determine the critical regions for each test, you will need to compute the

degrees of freedom separately for each test. For this problem, we are going to use an alpha of .05. The *F* table for an alpha of .05 is located in Appendix C.

3a. Define the Critical Region for the Interaction Test

The *df* of the numerator and the *df* of the denominator of the *F* value determine the critical value of every *F* test. It is easier to discuss *df*s if we introduce a bit of notation. Specifically, we will use *a* to indicate the number of levels of Factor A (studying method), *b* to indicate the number of levels of Factor B (time delay), and *N* to indicate the number of people in the study.

The numerator *df* for the interaction is computed as $df_{A \times B} = (a - 1)(b - 1)$. In this case, both factors have two levels, and so $df_{A \times B} = (2 - 1)(2 - 1) = 1$. The denominator *df* is computed as $df_{\text{within(error)}} = N - ab$. Thus, $df_{\text{within(error)}} = 32 - (2)(2) = 28$. If you look in the table of critical *F* values, you will find that the critical value of *F* with 1 and 28 degrees of freedom is 4.20.

Reading Question

25. If time delay had three levels (e.g., 5 minutes, 2 days, and 1 week) and study method had two levels, the degrees of freedom for the numerator of the interaction would be
    a. 1.
    b. 2.
    c. 3.

3b. Define the Critical Region for the First Main Effect Test (Factor A, Study Method)

The numerator *df* for the main effect of study method is computed as $df_A = (a - 1)$. Study method has two levels, so $df_A = 2 - 1 = 1$. The *df* for the denominator is the same as it is for the interaction: $df_{\text{within(error)}} = N - ab = 32 - (2)(2) = 28$. As before, the critical value of *F* with 1 and 28 degrees of freedom is 4.20.

Reading Question

26. If 40 people had participated in the study, the degrees of freedom for the denominator of the main effect of study method would be
    a. 40.
    b. 36.
    c. 42.

3c. Define the Critical Region for the Second Main Effect Test (Factor B, Delay Time)

The numerator *df* for the main effect of delay time is computed as $df_B = (b - 1)$. Delay time has two levels, so $df_B = 2 - 1 = 1$. The *df* for the denominator is the same as it is for the interaction: $df_{\text{within(error)}} = N - ab = 32 - (2)(2) = 28$. As before, the critical value of *F* with 1 and 28 degrees of freedom is 4.20.

Reading Question

27. If time delay had three levels (e.g., 5 minutes, 2 days, and 1 week), the degrees of freedom for the numerator of the main effect would be
    a. 1.
    b. 2.
    c. 3.

In this case, all three sets of *df*s were identical. While the denominator *df*s are always the same for all three tests, the numerator *df*s will *not* always be the same. When Factor A and Factor B have a different number of levels, the numerator *df*s will be different for the two main effect tests. Consequentially, when the factors have a different number of levels, the two main effect tests will have different critical values.

Step 4: Compute the Test Statistics (Three *F* Tests)

To determine if the three effects (i.e., the interaction and the two main effects) are statistically significant, you need to compute three *F* values. Although you could do these computations by hand, we will be doing these computations using SPSS. When you analyze the data in SPSS, you obtain an ANOVA summary table similar to the one displayed in Table 12.9. Table 12.10 includes the formulas used to

Table 12.9 ANOVA Source Table for a Two-Way Independent Samples ANOVA

Source	SS	df	MS	F
Between	39.375	3		
Studying method	3.125	1	3.125	0.94
Time delay	21.125	1	21.125	6.33
Studying Method × Time Delay	15.125	1	15.125	4.53
Within (error)	93.500	28	3.339	
Total	132.875	31		

Table 12.10 Formulas for *df*, *MS*, and *F* for a Two-Way Independent Samples ANOVA

Source	SS	df	MS	F
Between	39.375			
Studying method (A)	3.125	$a - 1$	$\frac{SS_A}{df_A}$	$\frac{MS_A}{MS_{within(error)}}$
Time delay (B)	21.125	$b - 1$	$\frac{SS_B}{df_B}$	$\frac{MS_B}{MS_{within(error)}}$
Studying Method × Time Delay (A × B)	15.125	$(a - 1)(b - 1)$	$\frac{SS_{A\times B}}{df_{A\times B}}$	$\frac{MS_{A\times B}}{MS_{within(error)}}$
Within (error)	93.500	$N - (a)(b)$	$\frac{SS_{within(error)}}{df_{within(error)}}$	
Total	132.875	$N - 1$		

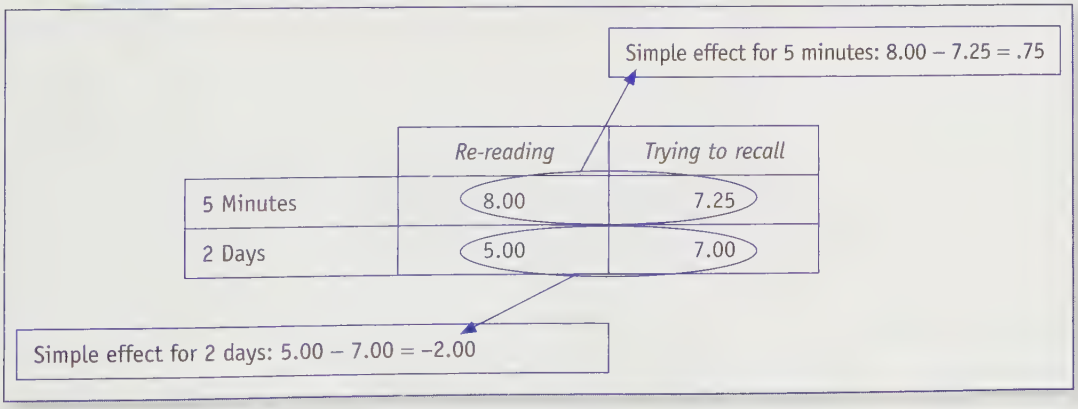
obtain the *df*, *MS* (mean squares), and *F*s. In future activities, we will work with the complete ANOVA summary table. For now, we will focus on interpreting the *F* values.

4a. F Test for the Interaction

There are three *F* values in Table 12.9. The *F* value testing the interaction between studying method and time delay is 4.53. Previously, you determined that the critical *F* value for the interaction *F* test is 4.20. The obtained *F* value of 4.53 is larger than the critical *F* value, so you reject the interaction null hypothesis. This means that the different pattern of simple effects for 5-minute and 2-day delays shown in Figure 12.1 are not likely to be due to sampling error. Specifically, this means that the difference between the simple effect for 5-minute delay, 8.00 – 7.25 = .75, is significantly different from the difference between the simple effect for 2-day delay, 5.00 – 7.00 = –2.00. So, finding a significant interaction means that the difference between .75 and –2.00 is unlikely to be created by sampling error.

Whenever you find a significant interaction, you will need to describe the simple effects in detail when you write your summary of the results. That means that you must know if the difference described by each simple effect is a significant difference. So, you need to know, if the test is 5 minutes later, whether the mean for rereading (8.00) is significantly higher than the mean for trying to recall (7.25). You also need to know, if the test is 2 days later, whether the mean for rereading (5.00) is significantly lower than the mean for trying to recall (7.00). Answering both of these questions requires additional significance tests called simple effects analyses. We will show you how to make SPSS perform these analyses later in the chapter. But, for now, we will tell you that the simple effect for 5 minutes (.75) was not a significant difference and the simple effect for 2 days was a significant difference (–2.00). You will need to know these simple effect details when you interpret any significant interaction.

Figure 12.1 Simple Effects Tested by Interaction *F* Test



4b. *F* Test for the First Main Effect (Study Method)

The *F* value for the main effect of study method is 0.94. The critical value of *F* for this main effect is 4.20. The obtained *F* is not greater than the critical value, so you do not reject the studying method

null hypothesis and you conclude that *overall*, the two studying methods lead to similar exam scores for those who reread ($M = 6.50$) and those who tried to recall ($M = 7.13$). You will interpret this main effect in more detail when you summarize the results below.

4c. F Test for the Second Main Effect (Time Delay)

The F value for the main effect of time delay is 4.529, and the critical value of F for this main effect is 4.20. The obtained F is greater than the critical value, so you reject the time delay null hypothesis and conclude that *overall*, the participants who experienced the longer time delay ($M = 6.00$) had significantly lower exam scores than those who took the exam 5 minutes after studying ($M = 7.63$). You will interpret this main effect in more detail when you summarize the results below.

Reading Question

28. Which of the three effects were significant? Choose all that apply.
    a. Main effect of studying method
    b. Main effect of time delay
    c. Interaction between study method and time delay

Table 12.11 General Guidelines for Interpreting η_p^2

η_p^2	*Estimated Size of the Effect*
Close to .01	Small
Close to .06	Medium
Close to .14	Large

Step 5: Compute the Effect Sizes

5a. Effect Size for the Interaction (Study Method by Time Delay)

After determining if each null hypothesis is rejected, the next step is computing an effect size for each test and interpreting them. Partial eta squared (η_p^2) measures the effect size of F values (Table 12.11).

We will start by computing the effect size for the interaction effect. The formula for computing the effect size of an interaction is given as follows:

$$\eta_p^2 = \frac{SS_{A \times B}}{SS_{A \times B} + SS_{within(error)}} = \frac{15.125}{15.125 + 93.5} = .139.$$

The $SS_{A \times B}$ and $SS_{within(error)}$ are both in the first column of the ANOVA summary table. The effect size for the interaction between studying method and time delay is .14, which is a large effect. It is important to note that partial eta squared is different from eta squared. The distinction between eta squared and partial eta squared is beyond the scope of this book. For now, you should know that partial eta squared is what is computed by SPSS. Using partial eta squared rather than the "classic" eta squared means that the eta-squared values from all three F tests can sometimes sum to more than 1.

Reading Question

29. When interpreting η_p^2, values close to _____ are considered large effect sizes.
    a. .01
    b. .06
    c. .14

Partial eta squared quantifies the overall effect of the interaction on the dependent variable. However, to interpret the interaction correctly, you need an effect size for each simple effect. Partial

eta squared does not give you this important information, so you must compute *d* for *each* simple effect difference. The *d* formula for simple effects is the same as the one you used with the one-way ANOVAs:

$$d = \frac{\textit{simple effect mean difference}}{SD_p},$$

where the pooled standard deviation is

$$SD_p = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{(n_1 - 1) + (n_2 - 1)}}.$$

In this study, you had two simple effects. The first was the effect of study method (reread vs recall) after a 5-minute time delay. People who studied by rereading the material (M = 8.00, SD = 2.00) had higher scores than people who studied by trying to recall the information (M = 7.25, SD = 1.91). The simple effect is 8.00 – 7.25 = .75. To convert this difference into an effect size (*d*), you divide the mean difference (the simple effect) by the pooled standard deviation of the two standard deviations from the two cells compared by the simple effect (SD_p^2). In this case, the pooled standard deviation is

$$SD_p = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{(n_1 - 1) + (n_2 - 1)}} = \sqrt{\frac{(7)2^2 + (7)1.91^2}{(7) + (7)}} = 1.956.$$

And, the effect size for this simple effect is

$$d = \frac{\textit{simple effect mean difference}}{SD_p} = \frac{8.00 - 7.25}{1.956} = .383.$$

The simple effect of study method after a 2-day delay is completed in the same way. After 2 days, people who studied by rereading the material (M = 5.00, SD = 1.69) had lower scores than people who studied by trying to recall the information (M = 7.00, SD = 1.69). The pooled standard deviation for this comparison is

$$SD_p = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{(n_1 - 1) + (n_2 - 1)}} = \sqrt{\frac{(7)1.69^2 + (7)1.69^2}{(7) + (7)}} = 1.69.$$

The effect size for this comparison is

$$d = \frac{\textit{simple effect mean difference}}{SD_p} = \frac{5.00 - 7.00}{1.69} = -1.18.$$

Note that the sign of the effect sizes indicates the direction of the effect. After a 5-minute delay (i.e., the first simple effect), the people who reread the material did better than those who tried to recall the material, yielding a positive effect size that is small to medium. After a 2-day delay (i.e., the second

simple effect), you subtracted in the same direction (reread – recall), which yielded a negative effect that is large.

5b. Effect Size for the First Main Effect (Studying Method)

The formula for computing the overall effect size for the main effect of studying method is similar to the formula for computing the overall effect size for the interaction:

$$\eta_p^2 = \frac{SS_{\text{studying method}}}{SS_{\text{studying method}} + SS_{\text{within (error)}}} = \frac{3.125}{3.125 + 93.5} = .032.$$

The effect size is 0.032, which is a small to medium effect.

When interpreting simple effects and all other pairwise comparisons, we have used d. Although d and partial eta squared are both measures of effect size that allow you to interpret the size of the effect, they are interpreted differently. Since we are computing d for the interaction, you can more easily compare effect sizes if you also compute d for the main effect, studying method. In this case, there were only two marginal means for studying method. People who reread the material had lower scores ($M = 6.50$, $SD = 2.37$, $n = 16$) than people who tried to recall the information ($M = 7.13$, $SD = 1.75$, $n = 16$). As earlier, you must first compute the pooled standard deviation and then the effect size. In this case, the pooled standard deviation is

$$SD_p = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{(n_1 - 1) + (n_2 - 1)}} = \sqrt{\frac{(15)2.37^2 + (15)1.75^2}{(15) + (15)}} = 2.083.$$

And, the effect size for this main effect is

$$d = \frac{\text{simple effect mean difference}}{SD_p} = \frac{6.50 - 7.13}{2.083} = -.302.$$

The effect size (d) is –.30, a small to medium effect, which is consistent with the partial eta squared results. In this case, there were only two marginal means. If you had more than two marginal means, you could compute d for each pairwise comparison.

5c. Effect Size for the Second Main Effect (Time Delay)

The formula for computing the effect size for the main effect of time delay is given as follows:

$$\eta_p^2 = \frac{SS_{\text{time delay}}}{SS_{\text{time delay}} + SS_{\text{within(error)}}} = \frac{21.125}{21.125 + 93.5} = .184.$$

The effect size is .184, which is a medium to large effect.

Again, you can also compute d for the marginal means. In this case, scores were higher after a 5-minute delay ($M = 7.63$, $SD = 1.93$, $n = 16$) than after a 2-day delay ($M = 6.00$, $SD = 1.93$, $n = 16$). The pooled standard deviation is

$$SD_p = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{(n_1 - 1) + (n_2 - 1)}} = \sqrt{\frac{(15)1.93^2 + (15)1.93^2}{(15) + (15)}} = 1.93.$$

$$d = \frac{\text{simple effect mean difference}}{SD_p} = \frac{7.63 - 6.00}{1.93} = .845.$$

The effect size (d) is .845, which is a large effect and is consistent with the partial eta squared results.

Reading Question

30. When computing the effect size for a two-way ANOVA,
    a. you will only need to compute one effect size value for all three F tests in a two-way ANOVA.
    b. you will have to compute separate effect sizes for each F test.

Reading Question

31. Which effect size describes the size of the simple effects when interpreting an interaction?
    a. η_p^2
    b. d

Step 6: Writing Up the Results of a Two-Way ANOVA

Students can find summarizing a two-way ANOVA daunting because there are three different F values and lots of statistical information that must be presented in a very specific format. However, there is a general format you can follow that should make writing the results easier.

General Format

1. Create a table to report the cell means and standard deviations as well as the marginal means and standard deviations. Once this information is in the table, you do not need to report the means and standard deviations in your write-up. Instead, refer the readers to the table.
2. Tell whether or not the interaction was significant, and report the statistical information in the correct format.
3. If the interaction is significant, describe the interaction, using one sentence to describe each simple effect. Include the p value from the simple effect as well as the d.
4. Tell whether or not the first main effect was significant, and report the statistical information in the correct format.
5. If the first main effect was significant, describe which marginal mean(s) was (were) significantly higher.
6. Tell whether or not the second main effect was significant, and report the statistical information in the correct format.
7. If the second main effect was significant, describe which marginal mean(s) was (were) significantly higher.

In this chapter, we did not do all of the computations by hand and so you still need additional information from SPSS to complete the APA-style write-up. The next section shows you how to use SPSS to obtain all of the necessary analyses. We will return to the write-up after the SPSS section.

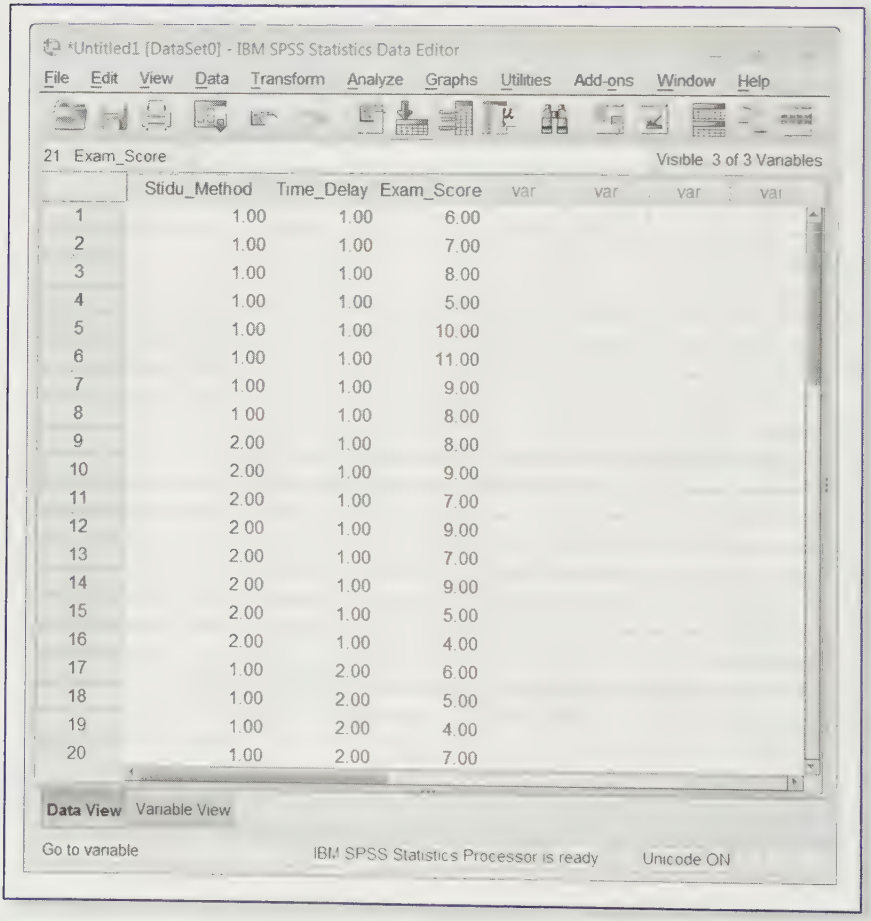
SPSS

The following pages illustrate how to use SPSS to analyze the same "studying method study" discussed throughout this chapter.

Data File

The SPSS data file should have three columns: one for the first IV (studying method), one for the second IV (time delay), and one for the DV (exam score). The studying method and time delay variables are used to indicate which group (i.e., cell) each participant was in. In this file, we used a "1" to indicate that the person reread and a "2" to indicate that the person recalled, but any two numbers can be used to label the two groups. For the time delay variable, we used a "1" to indicate that the person took the exam 5 minutes later and a "2" to indicate that the person took the exam 2 days later. Your file should have 32 rows, one row for each person. The first eight people in our data file have a "1" and a "1" followed by their DV score. The next eight people have a "2" and a "1" followed by their DV score. The third group of eight people have a "1" and a "2" followed by their DV score. The final group of eight people have a "2" and a "2" followed by their DV score. The first 15 lines can be seen in Figure 12.2.

Figure 12.2 SPSS Screenshot of Data Entry Screen



Reading Question

32. When entering the data for a two-way ANOVA, you will need two columns to indicate
    a. which combination of IV conditions each participant was in.
    b. each person's DV score.

To compute a two-factor ANOVA:

- Click on the Analyze menu. Choose General Linear Model, and then select Univariate.
- Move the DV into the Dependent Variable box.
- Move both IVs to the Fixed Factors box.
- Click on Options, and then Descriptive Statistics and Estimates of Effect Size.
- You can create *graphs* by clicking on the Plots button.
  - To graph the interaction, choose one IV (i.e., grouping variable) to put on the horizontal axis and one IV to make separate lines for. Click Add and then Continue. In this case, we put time delay on the horizontal axis and made separate lines for studying method.
  - To obtain *confidence intervals:*
  - Move everything that is in the Factor(s) and Factor interactions into the Display Means for box.
  - Click on Compare Main effects.
  - Select LSD(none) from the Confidence Interval Adjustment drop-down box.
  - Click on Continue and then OK.
  - We did not include the confidence interval output for this analysis below. The confidence intervals are interpreted just like those in previous chapters.

To compute simple effects analysis for a two-factor ANOVA:

- If you have a significant interaction, you need to rerun the ANOVA to obtain the simple effects analyses. Unfortunately, you cannot obtain simple effects tests through the point-and-click method, so you will need to learn a new procedure.
- Begin the same way as when you conducted the ANOVA. Click on the Analyze menu. Choose General Linear Model, and then select Univariate.
- Move the DV into the Dependent Variable box.
- Move both IVs (in this case, study method and time delay) to the Fixed Factors box.
- Click on Options and then
  - Move one IV (in this case, study method) into the Display Means for box
  - Click on Compare Main effects
  - Select Bonferroni from the Confidence Interval Adjustment drop-down box
  - Click on Continue
- Click the Paste button. This will open a syntax window. You need to change one line of syntax to get SPSS to run the simple effects analyses.
- In the syntax file, you will see something like the following:
  - /EMMEANS = TABLES(Study_Method) COMPARE ADJ(BONFERRONI)
  - This syntax needs to be edited so that it includes both IVs: /EMMEANS = TABLES(Study_Method*Time_Delay) COMPARE (Time_Delay) ADJ(BONFERRONI)
  - After you have edited this line, click on Run→All.
- After you run the analysis, much of the output will look like the output you obtained when doing the ANOVA. However, there is a new table that provides information about the simple effects. This table is titled "Pairwise comparisons" and is described at the end of the SPSS output.

Output File

Between-Subjects Factors

		Value Label	N
Study_Method	1.00	Re-reading	16
	2.00	Recalling	16
Time_Delay	1.00	5 minutes	16
	2.00	2 days	16

Between-Subjects Factors:

This table displays the IVs and codes used to enter the levels of each IV in the left-most box. The value labels for each numerical code are displayed in the "Value Label" box and the number of people in each IV level is displayed in the "N" box.

Descriptive Statistics

Dependent Variable: Exam_Score

Study_Method	Time_Delay	Mean	Std. Deviation	N
Re-reading	5 minutes	8.0000	2.00000	8
	2 days	5.0000	1.69031	8
	Total	6.5000	2.36643	16
Recalling	5 minutes	7.2500	1.90863	8
	2 days	7.0000	1.69031	8
	Total	7.1250	1.74642	16
Total	5 minutes	7.6250	1.92787	16
	2 days	6.0000	1.93218	16
	Total	6.8125	2.07034	32

Std. Deviation:

Displays the standard deviation (*SD*) for the mean on the same row

N:

Displays the sample size (*n*) for the mean on the same row

IV Conditions:

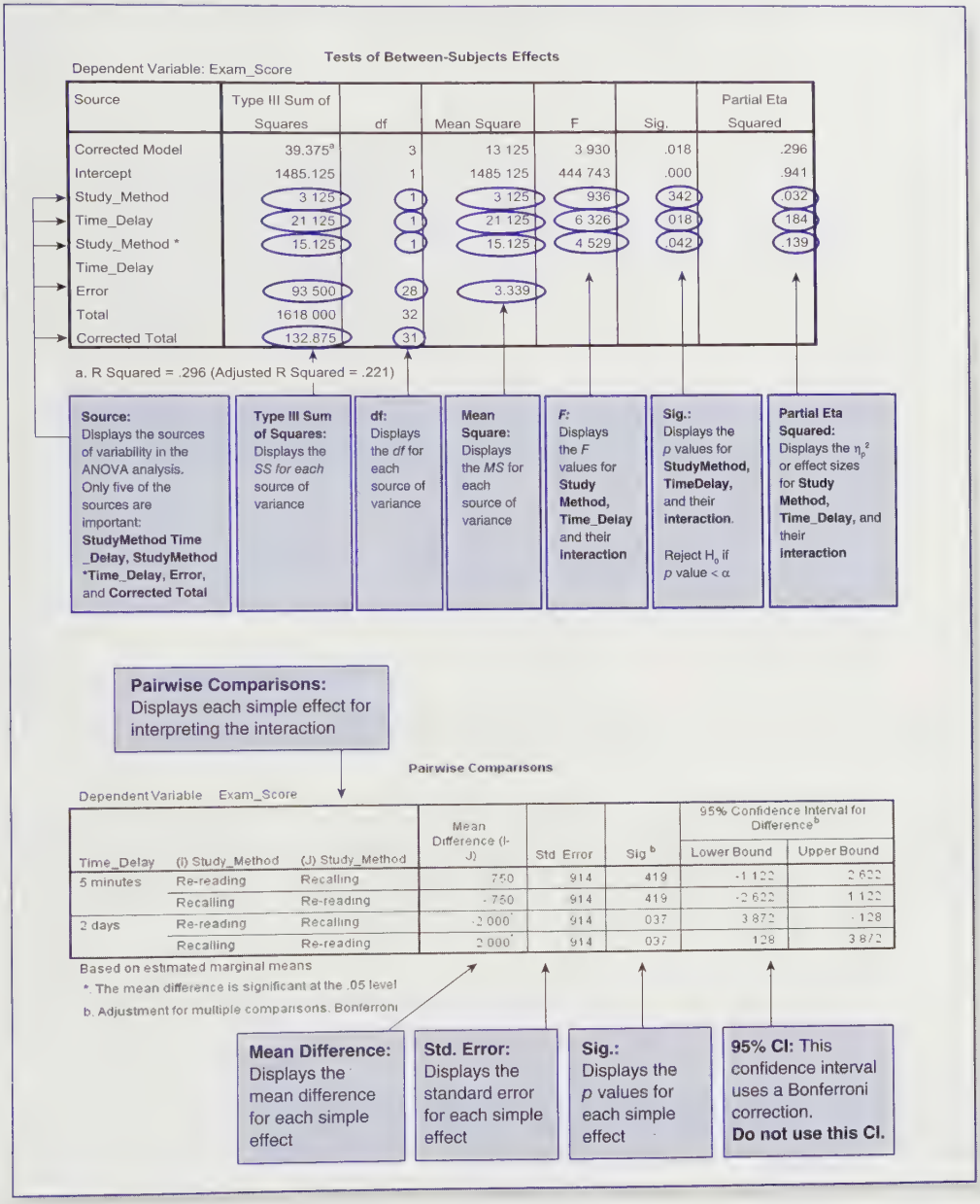
Each row displays statistical information for a cell in the factorial design or statistical information for a marginal mean.

Mean:

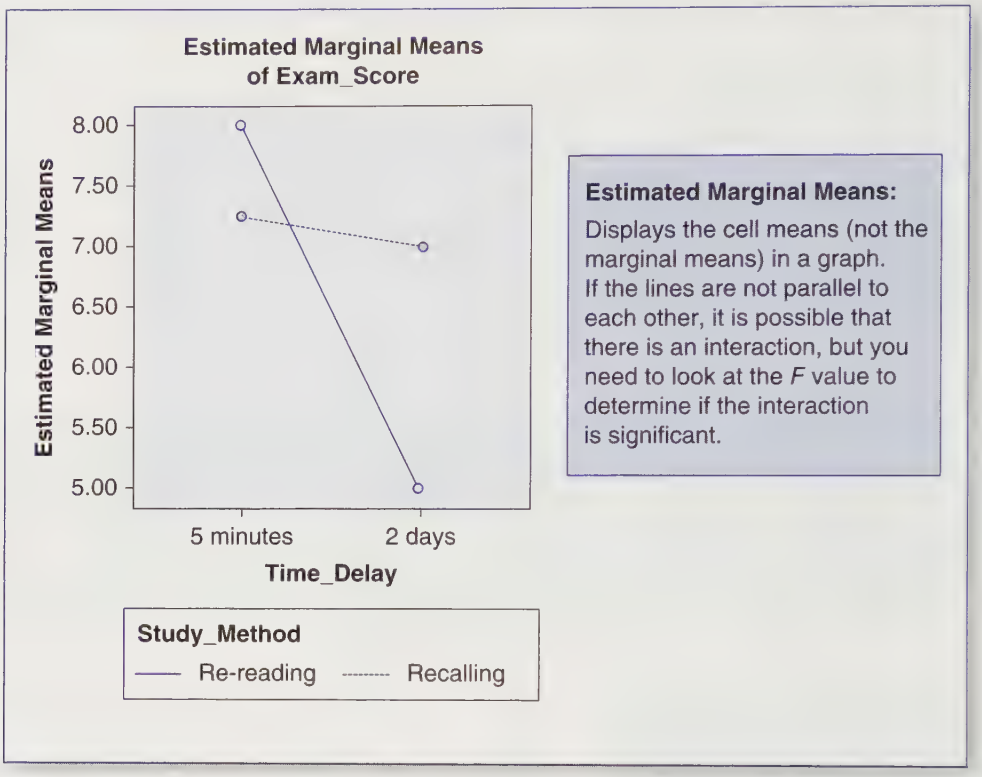
Each row displays a cell or marginal mean. The means are organized a bit differently. The following table is a more conventional way to organize these means:

	Rereading	Recalling	Marginal Mean
5 minute delay	8.0	7.25	7.625
2 day delay	5.0	7.0	6.0
Marginal Mean	6.5	7.125	Total = 10.5

Confirm that you can identify where each of the above means are located in the SPSS output table. You will need to use the labels on the left of the SPSS output table (IV Conditions) to help you identify the means.



The preceding "Pairwise Comparisons" output contains the simple effects analysis you need to interpret the significant interaction between studying method and time delay. The simple effect for a 5-minute time delay was not significant, indicating that, when the exam was taken 5 minutes after studying, there was no meaningful difference between the studying methods ($p = .419$). However, the simple effect for a 2-day delay was significant, indicating that, when the exam was taken 2 days after studying, the trying to recall strategy was better than the rereading strategy ($p = .037$).



Reading Question

33. Use the "Between-Subjects Factors" output to determine how many people reread as their studying method. According to the output, ____ people reread as their studying method.

a. 0
b. 1
c. 8
d. 16

Reading Question

34. Use the "Descriptive Statistics" output to determine the cell mean for the 5-minute delay, recall condition. The cell mean for this condition was _____.

a. 7
b. 7.25
c. 8
d. 5

Reading Question

35. Use the "Tests of Between-Subjects Effects" output to determine the p value for the interaction. The p value for the interaction between study method and time delay was _____.

a. .342

b. .018

c. .042

Reading Question

36. The line graph in the SPSS output is titled "Estimated Marginal Means." Which means are included in this graph? (Hint: Look at the explanation box next to the graph in the SPSS output above.).

a. The marginal means for both main effects

b. The cell means for the interaction

When reporting the statistical information in APA format, you will need to extract the needed numbers from the "Tests of Between-Subjects Effects" output. The correct format for a main effect is $F(df_{Time_Delay}, df_{Within(error)})$ = obtained F value, p value, MSE, η_p^2 value. Therefore, the correct way to write the statistical information for the main effect of time delay is as follows: $F(1, 28) = 6.34$, $p = .02$, $MSE = 3.34$, $\eta_p^2 = .18$.

Reading Question

37. Use the "Tests of Between-Subjects Effects" output to determine which of the following is the correct statistical information for the interaction between study method and time delay:

a. $F(1, 28) = 4.53$, $p = .04$, $MSE = 3.34$, $\eta_p^2 = .14$.

b. $F(1, 28) = 6.34$, $p = .02$, $MSE = 3.34$, $\eta_p^2 = .18$.

APA-Style Summary

Now that you have completed all of the analyses, you can summarize the results using APA style. Be sure that you know where each number came from.

Table 1. Memory Scores After a 5-Minute or 2-Day Delay for Two Different Study Methods

	Rereading		*Trying to Recall*		*Study Method Main Effect*	
Delay	***n***	***M* (*SD*)**	***n***	***M* (*SD*)**	***n***	***M* (*SD*)**
5 minutes	8	8.00 (2.00)	8	7.25 (1.91)	16	7.63 (1.93)
2 days	8	5.00 (1.69)	8	7.00 (1.69)	16	6.00 (1.93)
Drug main effect	16	6.50 (2.37)	16	7.13 (1.75)		

A 2 (Studying Method: Rereading or Trying to Recall) × 2 (Time Delay: 5 Minutes or 2 Days) factorial ANOVA revealed a significant interaction between studying method and time delay on exam score, $F(1, 28) = 4.53$, $p = .04$, $MSE = 3.54$, $\eta_p^2 = .14$ (see Table 1 for means and standard deviations). The results indicate that when taking an exam 5 minutes after studying, participants who studied by rereading versus recalling did not have meaningfully different exam scores, $p = .42$, $d = .37$. However, when taking an exam 2 days after studying, participants who studied by trying to recall material scored substantially higher than those who studied by rereading, $p = .04$, $d = 1.18$. There was also a significant main effect of time delay: $F(1, 28) = 6.33$, $p = .02$, $\eta_p^2 = .18$. Overall, participants who took the exam 5 minutes after studying scored significantly higher than those who took the exam 2 days later, $d = .84$. Finally, the main effect of studying method was not significant, $F(1, 28) = 0.94$, $p = .34$, $\eta_p^2 = .03$. When the results were averaged across time delay conditions, the participants who tried to recall the information did not receive significantly higher scores than the participants who reread the information ($d = -.31$). This main effect suggests that, overall, the studying methods were equally effective, but this main effect is misleading. The significant interaction suggests that when students are studying immediately before an exam, rereading and trying to recall are equally effective studying methods, but when studying for a more distant exam, trying to recall material is a far superior studying strategy.

Overview of the Activities

In Activities 12.1, 12.2, and 12.3, you will work to understand the logic of the two-factor ANOVA and learn to interpret the results in an APA-style write-up. In Activities 12.1 and 12.2, the SPSS output files are given to you, while in Activity 12.3, you will analyze some of the data in SPSS yourself. In Activity 12.4, you will review one-way independent measures ANOVAs and contrast them with two-way independent measures ANOVAs. In Activity 12.5, you will practice reading research scenarios and determining which statistic is appropriate to answer the research question.

Activity 12.1: Two-Factor ANOVAs I

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Identify the main effects and the interaction effect produced by a two-way ANOVA
- Identify the null hypotheses for both main effects and the interaction effect
- Create main effects and interaction graphs of a two-way ANOVA
- Clearly summarize the results of a two-factor ANOVA

In this activity, you will work with a series of simple research scenarios involving the effectiveness of two drugs (Drug A and Drug B) on memory for men and women. For the first scenario, much of the information is provided for you, but as you work through the activity, you will generate more of the information yourself. All of the two-factor ANOVAs in this activity make the same statistical assumptions—namely, *data independence*, *normality*, *homogeneity of variance*, and *appropriate measurement of the IV and the DV*. All of these assumptions were met by these data.

SCENARIO 1

You want to determine if Drug A or Drug B is more effective for *improving* memory in men and women. You gave a sample of 32 people (16 males and 16 females) either Drug A or Drug B, followed by a memory test.

1. This two-factor ANOVA created three significance tests. Each test has its own null and research hypothesis. Match each of the following statements to the null or research hypothesis it represents.

_____ Drug A and Drug B will be equally effective at increasing memory performance.
_____ Drug A and Drug B will not be equally effective at increasing memory performance.
_____ Men and women will have equally good memory performance.
_____ Men and women will not have equally good memory performance.
_____ The drugs will have similar effects on men and women.
_____ The drugs will not have similar effects on men and women.

a. Null hypothesis for interaction between drug type and gender
b. Research hypothesis for interaction between drug type and gender
c. Null hypothesis for main effect of drug type
d. Research hypothesis for main effect of drug type
e. Null hypothesis for main effect of gender
f. Research hypothesis for main effect of gender

The relevant portions of the SPSS output are given as follows:

Descriptive Statistics

Dependent Variable: MemoryScore

Gender	Drug	Mean	Std. Deviation	N
male	A	19.6250	2.32609	8
	B	16.7500	1.98206	8
	Total	18.1875	2.56174	16
female	A	17.1250	2.23207	8
	B	19.8750	2.23207	8
	Total	18.5000	2.58199	16
Total	A	18.3750	2.55278	16
	B	18.3125	2.60048	16
	Total	18.3438	2.53504	32

Tests of Between-Subjects Effects

Dependent Variable: MemoryScore

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	64.094[a]	3	21.365	4.427	.011	.322
Intercept	10767.781	1	10767.781	2231.252	.000	.988
Gender	.781	1	.781	.162	.690	.006
Drug	.031	1	.031	.006	.936	.000
Gender * Drug	63.281	1	63.281	13.113	.001	.319
Error	135.125	28	4.826			
Total	10967.000	32				
Corrected Total	199.219	31				

a. R Squared = .322 (Adjusted R Squared = .249)

Pairwise Comparisons

Dependent Variable: MemoryScore

Gender	(I) Drug	(J) Drug	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
						Lower Bound	Upper Bound
male	A	B	2.875*	1.098	.014	.625	5.125
	B	A	-2.875*	1.098	.014	-5.125	-.625
female	A	B	-2.750*	1.098	.018	-5.000	-.500
	B	A	2.750*	1.098	.018	.500	5.000

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

2. Insert the means, standard deviations, and sample sizes into the following table:

	Drug A		*Drug B*		*Gender Main Effect*
Gender	***n***	***M (SD)***	***n***	***M (SD)***	***M (SD)***
Males	_____	_____	_____	_____	_____
Females	_____	_____	_____	_____	_____
Drug main effect		_____		_____	

3. Record the statistics for the Gender × Drug interaction using APA style:

F(_____, _____) = _____, p = _____, MSE = _____, η_p^2 = _____.

In this study, the interaction was significant, indicating that there was a meaningful difference between the pattern of results for men and for women. To describe these differing results (i.e.,

to describe the interaction between drug and gender), you must compute the simple effects and then complete a simple effects analysis to determine which simple effect differences are significantly different.

4. Compute the mean difference for the simple effect of males across Drug A and Drug B.

   Simple effect: _____ – _____ = _____

5. Is the simple effect for males significant across Drug A and Drug B? (Look at the pairwise comparison output)
   a. Yes, the p value is less than .05.
   b. No, the p value is greater than .05.

6. Next you need to compute the effect size of the simple effect for males across Drug A and Drug B. You need the pooled standard deviation (SD_p) for this comparison. You can find the *SD*s you need in the provided SPSS output.

$$SD_p = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{(n_1 - 1) + (n_2 - 1)}} =$$

7. Now you can compute the effect size of the simple effect for males across Drug A and Drug B by dividing the simple effect mean difference by the pooled standard deviation you just computed.

$$d = \frac{\text{simple effect mean difference}}{SD_p} =$$

8. Compute the mean difference for the simple effect of females across Drug A and Drug B.

9. Is the simple effect for females across Drug A and Drug B significant? (Look at the pairwise comparison output.)
   a. Yes, the p value is less than .05.
   b. No, the p value is greater than .05.

10. Compute the effect size of the simple effect for females across Drug A and Drug B. You can find the *SD*s you need in the provided SPSS output.

$$SD_p = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{(n_1 - 1) + (n_2 - 1)}} =$$

$$d = \frac{\text{mean difference}}{SD_p} =$$

11. Which of the following is the *best* description of this interaction?
    a. Women had higher scores on the memory test than men. In addition, Drug A was more effective than Drug B.
    b. Women who took Drug B had significantly higher scores on the memory test than women who took Drug A. The pattern of results was the opposite for men. Men who took Drug B had significantly lower scores on the memory test than men who took Drug A.

12. Record the statistics for the gender main effect using APA style:

    F(_____, _____) = _____, p = _____, MSE = _____, η_p^2 = _____.

13. Record the means and standard deviations for the two means being compared by the F value above:

    Females (M = _____, SD = _____); *Males* (M = _____, SD = _____).

14. Is the main effect of gender significant?
    a. Yes, the p value is less than .05.
    b. No, the p value is greater than .05.

15. Compute the effect size (d) for the main effect of gender (i.e., comparing males and females).

$$SD_p = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{(n_1 - 1) + (n_2 - 1)}} =$$

$$d = \frac{\text{mean difference}}{SD_p} =$$

16. Which of the following is the best interpretation of the main effect of gender?
    a. Males had higher scores than did females.
    b. Females had higher scores than did males.
    c. Males' and females' scores were not significantly different.

17. Record the statistics for the drug main effect using APA style:

    F(_____, _____) = _____, p = _____, MSE = _____, η_p^2 = _____.

18. Record the means and the standard deviations for the two means being compared by the F value above:

    Drug A (M = _____, SD = _____); Drug B (M = _____, SD = _____).

19. Is the main effect of drug significant?
    a. Yes, the p value is less than .05.
    b. No, the p value is greater than .05.

20. Compute d to compare the main effect of drug: Drug A vs. Drug B.

$$SD_p = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{(n_1 - 1) + (n_2 - 1)}} =$$

$$d = \frac{\text{mean difference}}{SD_p} =$$

21. Which of the following is the best interpretation of the main effect of drug?
    a. Drug A was more effective than Drug B.
    b. Drug B was more effective than Drug A.
    c. The memory scores for Drugs A and B were not significantly different.

22. Questions 1 to 21 include all of the information necessary for constructing an APA-style write-up of these analyses. An example of a write-up is included below so that you can see how each part relates to the overall summary. Fill in the missing numbers.

Table 1. Memory Scores for Drugs A and B for Males and Females

	Drug A		*Drug B*		*Gender Main Effect*
Gender	***n***	***M (SD)***	***n***	***M (SD)***	***M (SD)***
Males	8	19.63 (2.33)	__8__	_____	18.19 (2.56)
Females	8	17.13 (2.23)	_____	19.88 (2.23)	_____
Drug main effect		18.38 (2.55)		_____	

There was a significant interaction between gender and drug treatment, $F(1, 28) = 13.11$, $p = .001$, $MSE = 4.83$, $\eta_p^2 = .32$. Simple effects analyses revealed that women who took Drug A had memory scores that were significantly lower than the memory scores for women who took Drug B, $p = .02$, $d = -1.23$. The pattern of results was the opposite for the men's simple effect. Men who took Drug A had memory scores that were significantly _____than did men who took Drug B, $p = .01$, $d =$ _____.

The main effect of gender was not statistically significant, $F(1,$ _____$) = 0.16$, $p =$ _____, $MSE =$ _____, $\eta_p^2 = .01$. Overall, males' memory scores were not significantly different from females' memory scores, $d =$ _____.

Finally, the main effect of drug was not statistically significant, $F(1, 28) = 0.01$, $p =$ _____, $MSE = 4.83$, $\eta_p^2 = .00$. Overall, memory scores for people taking Drug A were not significantly different from memory scores for people taking Drug B, $d =$ _____.

SCENARIO 2

Now, you want to determine if Drug A or Drug B is more effective for improving memory in elderly men and women, those older than 80 years. A sample of 32 elderly people (16 males and 16 females) were given either Drug A or Drug B, followed by a memory test. The SPSS output is given as follows:

Descriptive Statistics

Dependent Variable: MemoryScore

Gender	Drug	Mean	Std. Deviation	N
male	A	19.6250	2.32609	8
	B	15.7500	1.98206	8
	Total	17.6875	2.89180	16
female	A	20.3750	1.92261	8
	B	19.8750	2.23207	8
	Total	20.1250	2.02896	16
Total	A	20.0000	2.09762	16
	B	17.8125	2.94887	16
	Total	18.9063	2.75165	32

Tests of Between-Subjects Effects

Dependent Variable: MemoryScore

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	108.594[a]	3	36.198	8.036	.001	.463
Intercept	11438.281	1	11438.281	2539.321	.000	.989
Gender	47.531	1	47.531	10.552	.003	.274
Drug	38.281	1	38.281	8.499	.007	.233
Gender * Drug	22.781	1	22.781	5.057	.033	.153
Error	126.125	28	4.504			
Total	11673.000	32				
Corrected Total	234.719	31				

a. R Squared = .463 (Adjusted R Squared = .405)

Pairwise Comparisons

Dependent Variable: MemoryScore

Gender	(I) Drug	(J) Drug	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
						Lower Bound	Upper Bound
male	A	B	3.875*	1.061	.001	1.701	6.049
	B	A	-3.875*	1.061	.001	-6.049	-1.701
female	A	B	.500	1.061	.641	-1.674	2.674
	B	A	-.500	1.061	.641	-2.674	1.674

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

23. Insert the means, standard deviations, and sample sizes into the following table:

	Drug A		*Drug B*		*Gender Main Effect*
Gender	***n***	***M (SD)***	***n***	***M (SD)***	***M (SD)***
Males	_____	_____	_____	_____	_____
Females	_____	_____	_____	_____	_____
Drug main effect		_____		_____	

24. Record the statistics for the Gender × Drug interaction using APA style:

F(_____, _____) = _____, p = _____, MSE = _____, η_p^2 = _____.

25. Compute the mean difference for the simple effect of males across Drug A and Drug B.

26. Was the simple effect difference for males across Drugs A and B statistically significant?
    a. Yes
    b. No

27. Compute the effect size (d) for the male simple effect across Drugs A and B. (Hint: You need two equations.)

28. Compute the mean difference for the simple effect of females across Drug A and Drug B.

29. Was the simple effect difference for females across Drugs A and B statistically significant?
    a. Yes
    b. No

30. Compute the effect size (d) for the female simple effect across Drugs A and B. (Hint: You need two equations.)

31. Which of the following is the *best* description of this interaction?
    a. Both men and women who took Drug A had higher scores on the memory test than did those who took Drug B.
    b. Women who took Drug A had higher scores on the memory test than did women who took Drug B. Men who took Drug A had higher scores on the memory test than did men who took Drug B.
    c. Women who took Drug A had scores on the memory test that were not significantly different from the scores for women who took Drug B. Men who took Drug A had scores on the memory test that were significantly higher than those for men who took Drug B.

32. Record the statistics for the gender main effect using APA style:

F(_____, _____) = _____, p = _____, MSE = _____, η_p^2 = _____.

33. Record the means and standard deviations for the two means being compared by the F value above:

Females (M = _____, SD = _____); Males (M = _____, SD = _____).

34. Which of the following is the best interpretation of the main effect of gender?
    a. Males had significantly higher scores than females.
    b. Females had significantly higher scores than males.
    c. Males' and females' scores were not significantly different.

35. Compute the effect size (d) to compare males and females for the main effect of gender. (Hint: You need two equations.)

36. Record the statistics for the drug main effect using APA style:

F(_____, _____) = _____, p = _____, MSE = _____, η_p^2 = _____.

37. Record the means and standard deviations for the two means being compared by the F value above:

Drug A (M = _____, SD = _____); Drug B (M = _____, SD = _____).

38. Which of the following is the best interpretation of the main effect of drug?
    a. Drug A was significantly more effective than Drug B.
    b. Drug B was significantly more effective than Drug A.
    c. The memory scores for Drugs A and B were not significantly different.

39. Compute the effect size (d) to compare Drugs A and B for the main effect of drug. (Hint: You need two equations.)

40. The APA-style summary statement below has *three* errors (one in each paragraph). Find the errors and fix them. The errors may be small, such as a missing number, or large, such as a missing sentence, a sentence that must be edited to include missing information, or a sentence that must be deleted.

Table 1. Memory Scores for Drugs A and B for Males and Females

	Drug A		*Drug B*		*Gender Main Effect*
Gender	***n***	***M* (*SD*)**	***n***	***M* (*SD*)**	***M* (*SD*)**
Males	8	19.63 (2.33)	8	15.75 (1.98)	17.69 (2.89)
Females	8	20.38 (1.92)	8	19.88 (2.23)	20.13 (2.03)
Drug main effect		20.00 (2.10)		17.81 (2.95)	

A two-factor ANOVA revealed a significant interaction between gender and drug treatment, $F(1, 28) = 5.06$, $p = .03$, $MSE = 4.50$, $\eta_p^2 = .15$. For elderly women, there was no significant difference in memory scores between those who took Drug A and those who took Drug B, $p = .64$, $d = .24$. However, for elderly men, those who took Drug A had scores on the memory test that were different from those who took Drug B, $p = .001$, $d = 1.80$.

The main effect of gender was also statistically significant, $F(1, 28) = 10.55$, $p = .003$, $MSE = 4.50$, $\eta_p^2 = .27$.

Finally, the main effect of drug was not significant, $F(1, 28) = 8.50$, $p = .007$, $MSE = 4.50$, $\eta_p^2 = .23$. Overall, Drug A was more effective than Drug B, $d = .86$.

SCENARIO 3

Now, you want to determine if Drug A or Drug B is more effective for improving memory in men and women between the ages of 18 and 25, or "young adults." A sample of 32 young adults (16 males and 16 females) were given either Drug A or Drug B, followed by a memory test. The SPSS output is given as follows:

Descriptive Statistics

Dependent Variable: MemoryScore

Gender	Drug	Mean	Std. Deviation	N
male	A	19.5000	2.44949	8
	B	15.3750	1.68502	8
	Total	17.4375	2.94321	16
female	A	20.3750	1.92261	8
	B	17.0000	2.00000	8
	Total	18.6875	2.57472	16
Total	A	19.9375	2.17466	16
	B	16.1875	1.97379	16
	Total	18.0625	2.79328	32

Tests of Between-Subjects Effects

Dependent Variable: MemoryScore

Source	Type III Sum of Squares	df	Mean Square	F	Sig	Partial Eta Squared
Corrected Model	126.125[a]	3	42.042	10.170	.000	.521
Intercept	10440.125	1	10440.125	2525.473	.000	989
Gender	12.500	1	12.500	3.024	.093	097
Drug	112.500	1	112.500	27.214	.000	493
Gender * Drug	1.125	1	1.125	.272	.606	010
Error	115.750	28	4.134			
Total	10682.000	32				
Corrected Total	241.875	31				

a. R Squared = .521 (Adjusted R Squared = .470)

Pairwise Comparisons

Dependent Variable: MemoryScore

Gender	(I) Drug	(J) Drug	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
						Lower Bound	Upper Bound
male	A	B	4.125*	1.017	.000	2.043	6.207
	B	A	-4.125*	1.017	.000	-6.207	-2.043
female	A	B	3.375*	1.017	.003	1.293	5.457
	B	A	-3.375*	1.017	.003	-5.457	-1.293

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni

41. Insert the means, standard deviations, and sample sizes into the following table:

	Drug A		*Drug B*		*Gender Main Effect*
Gender	***n***	***M (SD)***	***n***	***M (SD)***	***M (SD)***
Males	_____	_____	_____	_____	_____
Females	_____	_____	_____	_____	_____
Drug main effect		_____		_____	

42. Record the statistics for the Gender × Drug interaction using APA style:

$F($_____, _____$) =$ _____, $p =$ _____, $MSE =$ _____, $\eta_p^2 =$ _____.

43. Compute the mean difference for the simple effect of males across Drug A and Drug B.

44. Is the simple effect for males significant?

a. Yes

b. No

45. Compute the effect size for the male simple effect across Drugs A and B.

46. Compute the mean difference for the simple effect of females across Drug A and Drug B.

47. Is the simple effect for females significant?
    a. Yes
    b. No

48. Compute the effect size for the female simple effect across Drugs A and B.

49. Which of the following is the *best* description of this interaction?
    a. Women had significantly higher scores on the memory test than men. In addition, Drug A was more effective than Drug B.
    b. Women who took Drug B had significantly higher scores on the memory test than women who took Drug A. The pattern of results was the opposite for men. Men who took Drug B had significantly lower scores on the memory test than men who took Drug A.
    c. Drug A and Drug B affected the memory scores of males and females in similar ways (i.e., there was no interaction between drug type and gender).

50. Record the statistics for the gender main effect using APA style:

    F(_____, _____) = _____, p = _____, MSE = _____, η_p^2 = _____.

51. Record the means and standard deviations for the two means being compared by the F value above:

    Females (M = _____, SD = _____); Males (M = _____, SD = _____).

52. Compute the d for the comparison between females and males for the main effect of gender.

53. Which of the following is the best interpretation of the main effect of gender?
    a. Males had significantly higher scores than females.
    b. Females had significantly higher scores than males.
    c. Males' and females' scores were not significantly different.

54. Record the statistics for the drug main effect using APA style:

    F(_____, _____) = _____, p = _____, MSE = _____, η_p^2 = _____.

55. Record the means and standard deviations for the two means being compared by the F value above:

    Drug A (M = _____, SD = _____); Drug B (M = _____, SD = _____).

56. Compute the d for the comparison between Drugs A and B for the main effect of drug.

57. Which of the following is the best interpretation of the main effect of drug?

a. Drug A was significantly more effective than Drug B.
b. Drug B was significantly more effective than Drug A.
c. The memory scores for Drugs A and B were not significantly different.

58. Complete the APA-style summary below. *You may need to add numbers as well as add and/or delete sentences to fix this write-up.*

	Drug A		*Drug B*		*Gender Main Effect*
Gender	***n***	***M (SD)***	***n***	***M (SD)***	***M (SD)***
Males	_____	19.50 (2.45)	8	15.38 (1.69)	_____
Females	_____	20.39 (1.92)	8	_____	18.69 (2.57)
Drug main effect		19.94 (2.17)		_____	

The interaction between gender and drug treatment was not significant, $F(1, 28) = .27$, $p = .61$, $MSE = 4.13$, $\eta_p^2 = .01$. Drug A improved the memory scores of males more than it improved the memory scores of females.

The main effect of gender was not significant, $F(1, 28) = 3.02$, $p = .09$, $MSE = 4.13$, $\eta_p^2 = .10$. Overall, memory scores were not significantly different for males and females, $d =$ _____.

The main effect of drug was significant, $F(1, 28) = 27.21$, $p < .001$, $MSE = 4.13$, $\eta_p^2 = .49$.

Activity 12.2: Two-Factor ANOVAs II

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Read a description of an experimental study and identify the IV, its levels, and the DV
- Identify the null and research hypotheses for a specific factorial design
- Determine if each null hypothesis should be rejected
- Write a summary of the results of a factorial design

EXAMPLE BASED ON ACTUAL RESEARCH

The following scenario is based on an actual studies conducted by Langer, Holzner, Magnet, and Kopp (2005), and the data are consistent with their actual results. Langer and colleagues were interested in

assessing the effects of talking on a cell phone while driving. While a great deal of research evidence indicates that talking on a cell phone degrades one's ability to drive, many people either are unaware of this evidence or don't find the evidence sufficiently compelling to stop them from using their cell phone while driving. Most everyone recognizes that driving after drinking alcohol is a *very* bad idea. How does the driving deficit created by talking a cell phone compare to the driving deficit created by drinking alcohol?

There were 40 participants in the following study. All of them operated a high-fidelity driving simulator while either talking on their own cell phone or after consuming a "low dose" of alcohol (i.e., 4–5 g per 100 ml blood). Additionally, half the participants in each of these conditions had "high driving experience" (i.e., more than 50,000 km), and half had "low driving experience" (i.e., less than 20,000 km). The DV was the time it took the participants to step on the simulator's brake pedal when needed (called reaction time). It is worth pointing out that if a driver's reaction time slows by even a fraction of a second, it could result in a deadly accident. Therefore, any driving condition that significantly slows drivers' reaction times is potentially dangerous. Obviously, longer reaction times indicate *worse* driving performance.

1. What are the IVs? What are the levels of each IV?

   IV 1: Driving condition: Level 1 = _____, Level 2 = _____.

   IV 2: Driving experience: Level 1 = _____, Level 2 = _____.

Null and Research Hypotheses

2. Match each of the following statements to the null or research hypothesis it represents.

   a. The reaction times of those with high and low driving experience will be affected by cell phone use and alcohol in similar ways.

   b. The reaction times of those with high and low driving experience will be affected by cell phone use and alcohol differently.

   c. The reaction times of those who drink alcohol and those who talk on the cell phone while driving will be similar.

   d. The reaction times of those who drink alcohol and those who talk on the cell phone while driving will be different.

   e. The reaction times of those with high and low driving experience will be similar.

   f. The reaction times of those with high and low driving experience will be different.

   ____ Null hypothesis for the interaction between the IVs

   ____ Research hypothesis for the interaction between the IVs

   ____ Null hypothesis for the main effect of driving condition

   ____ Research hypothesis for the main effect of driving condition

   ____ Null hypothesis for the main effect of driving experience

   ____ Research hypothesis for the main effect of driving experience

   The data for this study are in the file "TWOWAY ANOVA cell phones.sav." The relevant SPSS output from this study is provided on page 478. Above each table, there is a short description of the statistical information that table provides. Try to run the analysis in SPSS. Your output should have the following tables. If your output does not have all of the following tables, review how to run a factorial ANOVA with simple effects analysis on p. 459.

This table provides the cell and marginal means for the two-way ANOVA.

Descriptive Statistics

Dependent Variable: Reaction_Time

Driving_Experience	Driving_Condition	Mean	Std. Deviation	N
low	Alcohol	1.2700	.15670	10
	Cell Phone	1.2000	.18257	10
	Total	1.2350	.16944	20
high	Alcohol	1.0050	.17232	10
	Cell Phone	1.1950	.17865	10
	Total	1.1000	.19668	20
Total	Alcohol	1.1375	.21018	20
	Cell Phone	1.1975	.17583	20
	Total	1.1675	.19367	40

This table is the ANOVA source table. It provides all 3 obtained F values, p values, and η_p^2.

Tests of Between-Subjects Effects

Dependent Variable: Reaction_Time

Source	Type III Sum of Squares	df	Mean Square	F	Sig	Partial Eta Squared
Corrected Model	.387[a]	3	.129	4.321	011	265
Intercept	54.522	1	54.522	1825.013	000	981
Driving_Experience	.182	1	.182	6.100	018	145
Driving_Condition	.036	1	.036	1.205	280	032
Driving_Experience * Driving_Condition	.169	1	.169	5.657	023	136
Error	1.076	36	.030			
Total	55.985	40				
Corrected Total	1.463	39				

a. R Squared = .265 (Adjusted R Squared = .203)

This table includes the simple effects analyses.

Pairwise Comparisons

Dependent Variable: Reaction_Time

Driving_Experience	(I) Driving_Condition	(J) Driving_Condition	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
						Lower Bound	Upper Bound
low	Alcohol	Cell Phone	.070	.077	.371	-.087	.227
	Cell Phone	Alcohol	-.070	.077	.371	-.227	.087
high	Alcohol	Cell Phone	-.190*	.077	.019	-.347	-.033
	Cell Phone	Alcohol	.190*	.077	.019	.033	.347

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

3. Graph the interaction between driving condition and driving experience. Put driving condition on the x-axis.

4. An interaction tests if the effect of one IV *depends* on the level of the other IV. For this example, an interaction would test if the effect of driving condition (cell phone vs. alcohol) *depends* on whether the participants had low or high driving experience. Thus, the interaction is testing whether the mean differences between the cell phone and alcohol conditions are different for drivers with low experience versus high experience.
   a. What is the mean difference for the simple effect of low driving experience across the cell phone and alcohol conditions?

      _____ – _____ = _____

   b. What is the mean difference for the simple effect of high driving experience across the cell phone and alcohol conditions?

      _____ – _____ = _____

5. The interaction is testing to see if these two simple effects are significantly different. If they are, there
   a. is a significant interaction.
   b. is not a significant interaction.

6. Is the interaction between driving condition and driving experience significant (i.e., unlikely to be due to sampling error)?
   a. Yes ($p = .023$)
   b. No ($p = .136$)

7. Use the above SPSS output to compute the effect size (d) for the low driving experience simple effect.

$$SD_p = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{(n_1 - 1) + (n_2 - 1)}} =$$

$$d = \frac{\text{mean difference}}{SD_p} =$$

8. Use the SPSS output on page 478 to compute the effect size (d) for the high driving experience simple effect.

9. Graph the main effect of driving experience.

10. Is the main effect of driving experience significant?
    a. Yes ($p < .05$)
    b. No ($p > .05$)

11. Use the above SPSS output to compute the effect size (d) for the main effect of driving experience.

12. Graph the main effect of driving condition.

13. Is the main effect of driving condition (cell phone vs. alcohol) significant?
    a. Yes
    b. No

14. Use the above SPSS output to compute the effect size (d) for the main effect of driving condition.

Reporting Results

15. Use the provided SPSS output and the effect sizes you calculated to write an APA style summary of these results. The general format you should follow is provided below. You can also look at the example on p. 476 of the chapter. Use a separate piece of paper to write your summary.

General Format

1. Create a table to report the cell means and standard deviations as well as the marginal means and standard deviations. Once this information is in the table, you do not need to report the means and standard deviations in your write-up. Instead, refer the readers to the table.
2. Tell if the interaction was significant, and report the statistical information in the correct format.
3. If the interaction is significant, describe the interaction, using one sentence each to describe the pattern of simple effects for each level of an IV.
4. Tell if the first main effect was significant, and report the statistical information in the correct format.

5. If the first main effect was significant, describe which marginal mean(s) was (were) significantly higher.
6. Tell if the second main effect was significant, and report the statistical information in the correct format.
7. If the second main effect was significant, describe which marginal mean(s) was (were) significantly higher.

Activity 12.3: Two-Factor ANOVAs III

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Identify the main effects and interaction effect produced by a two-way ANOVA
- Identify the null hypotheses for both main effects and the interaction effect
- Use SPSS to compute and then interpret a two-way independent ANOVA
- Create main effects and interaction graphs of a two-way ANOVA
- Clearly summarize the results of a two-factor ANOVA for a lay audience

SCENARIO 1

Research has shown that both exercise and meditation are effective in the treatment of depression and creating a more positive outlook on life (e.g., Josefsson, Lindwall, & Archer, 2014). You wonder if combining these two activities may be associated with greater optimism. To test this hypothesis, you recruit people to participate in a study based on whether they currently exercise regularly (at least four times a week for at least 30 minutes a day) and whether they currently meditate regularly (at least four times a week for at least 30 minutes a day). In total, you find 10 people who are in each of the following conditions: regular meditation/regular exercise, regular meditation/no exercise, no meditation/regular exercise, and no meditation/no exercise. You ask each participant to complete a survey that measures optimism using an interval/ratio scale where higher numbers indicate greater optimism. Previous research suggests that scores on the scale are normally distributed in the population. The means and standard deviations are included in the following table.

		Meditation	
		Regular Meditation	*No Meditation*
Exercise	Regular exercise	6.90 (2.33) $n = 10$	3.50 (1.58) $n = 10$
	No exercise	4.00 (1.41) $n = 10$	3.10 (2.03) $n = 10$

1. Select the two grouping variables (IVs).
   a. Exercise vs. no exercise
   b. Meditation vs. no meditation
   c. Optimism
2. What is the dependent variable?
   a. Exercise vs. no exercise
   b. Meditation vs. no meditation
   c. Optimism
3. This study has two grouping variables, and the DV is measured on an interval/ratio scale. What statistical assumption is satisfied by these facts?
   a. Normality
   b. Appropriate measurement of the variables
   c. Homogeneity of variance
   d. Data independence
4. None of the standard deviations in the table above are more than double any of the other standard deviations. What statistical assumption is satisfied by this fact?
   a. Normality
   b. Appropriate measurement of the variables
   c. Homogeneity of variance
   d. Data independence
5. The sample sizes are less than 30 in each cell. What statistical assumption might be violated if the scores are not normally distributed in the population?
   a. Normality
   b. Appropriate measurement of the variables
   c. Homogeneity of variance
   d. Data independence
6. Each of the three significance tests produced by two-factor designs has its own null and research hypothesis. For each of the significance tests produced by this 2 × 2 design, match the correct null and research hypothesis.

   Null hypothesis (H_0) for the *main effect of exercise*: _____.

   Research hypothesis (H_1) for the *main effect of exercise*: _____.

   Null hypothesis (H_0) for the *main effect of meditation*: _____.

   Research hypothesis (H_1) for the *main effect of meditation*: _____.

   Null hypothesis (H_0) for the *interaction* between exercise and meditation: _____.

   Research hypothesis (H_1) for the *interaction* between exercise and meditation: _____.

   a. The effect of meditation will be *the same* for people who do and do not exercise.
   b. The effect of meditation will be *different* for people who do and do not exercise.

c. The exercise and no-exercise groups *will not* differ in their responses to the positive outlook scale.

d. The exercise and no-exercise groups *will* differ in their responses to the positive outlook scale.

e. The meditation and no-meditation groups *will not* differ in their responses to the positive outlook scale.

f. The meditation and no-meditation groups *will* differ in their responses to the positive outlook scale.

SPSS Data Entry

The positive life outlook scores for each person are listed as follows:

Regular mediation and regular exercise:	9, 6, 7, 6, 3, 8, 7, 8, 4, 11
No meditation and regular exercise:	3, 5, 4, 6, 2, 1, 4, 3, 2, 5
Regular meditation and no exercise:	3, 6, 4, 4, 5, 3, 1, 4, 5, 5
No meditation and no exercise:	2, 3, 4, 3, 1, 6, 7, 2, 1, 2

Enter these data into SPSS. Be sure that you create three columns: one for each IV (exercise and meditation) and one for the DV (positive outlook). You will have to use a coding system for each IV. For example, for exercise, you might enter a "1" for regular exercise and a "0" for no exercise. For meditation, you might enter "1" for regular meditation and "0" for no meditation.

Value Labels

It would also be a good idea to create value labels, using the following steps:

- Click on the Variable View tab at the bottom of your screen.
- Click on the cell across from the Exercise_Group variable in the Values column.
- Enter a "1" in the Value box, and type a label, such as "Exercise," in the Label box. Click the Add button.
- Enter a "0" in the Value box, and type a label, such as "No Exercise," in the Label box. Click Add and then OK.
- Repeat this procedure for the Meditation_Group variable.

Running a Two-Way ANOVA

Run a two-way ANOVA by doing the following:

- Click on the Analyze menu. Choose General Linear Model, and then select Univariate.
- Move the DV into the Dependent Variable box (in this case, Positive_Outlook).
- Move both IVs (in this case, Exercise_Group and Meditation_Group) to the Fixed Factors box.
- You can create graphs by clicking on the Plots button.

- To graph the interaction, choose one IV (i.e., grouping variable) to put on the horizontal axis and one IV to make separate lines for. Click Add and then Continue. In this case, we put Exercise_Group on the horizontal axis and made separate lines for Meditation_Group.

- Click on Options, and then click on Descriptive Statistics and Estimates of Effect Size.

Interpreting the Interaction

7. Find the cell and marginal means in the SPSS output, and record them below:

	Meditation		*No Meditation*		*Exercise Group Main Effect*
	n	***M (SD)***	***n***	***M (SD)***	***M (SD)***
Exercise	_____	_____	_____	_____	_____
No exercise	_____	_____	_____	_____	_____
Meditation group main effect		_____		_____	

8. Graph the interaction of exercise group and meditation group. You already created this graph in SPSS, but you should also be able to create the graph by hand using the number in the table above. Either a line graph or a bar graph is acceptable.

9. Fill in the numbers for the results of the interaction:

   $F(___, ___) = _____$, $p = _____$, $MSE = _____$, $\eta_p^2 = _____$.

10. Is the interaction between exercise group and meditation group statistically significant?
    a. Yes
    b. No

Because the interaction is significant, you need to compute the simple effects analyses using SPSS.

Running Simple Effects Analyses for a Two-Way ANOVA

The first three steps are the same as when you run the ANOVA:

- Click on the Analyze menu. Choose General Linear Model, and then select Univariate.
- Move the DV into the Dependent Variable box (in this case, Positive_Outlook).
- Move both IVs (in this case, Exercise_Group and Meditation_Group) to the Fixed Factors box.

To obtain the simple effects:

- Click on Options.
- Move one IV (in this case, Exercise_Group) into the Display Means for box.
- Click on Compare Main effects.
- Select Bonferroni from the Confidence Interval Adjustment drop-down box.
- Click on Continue.
- Click the Paste button. This will open a syntax window. You need to change one line of syntax to get SPSS to run the simple effects analyses.
- In the syntax file you will see the following:
  - /EMMEANS = TABLES(Exercise_Group) COMPARE ADJ(BONFERRONI)
  - This syntax needs to be edited so that it includes both IVs: /EMMEANS = TABLES(Exercise_Group *Meditation_Group) COMPARE (Meditation_Group) ADJ(BONFERRONI)
  - After you have edited this line, click on Run→All.

11. Compute the mean difference for the simple effect of the exercise group.

12. Is this simple effect significant? (Look at the pairwise comparisons table in the SPSS output.)
    a. Yes
    b. No

13. Compute the effect size (*d*) for this simple effect.

14. Next, compute the mean difference for the simple effect of the no-exercise group.

15. Is this simple effect significant?
    a. Yes
    b. No

16. Compute the effect size (*d*) for this simple effect for the no-exercise group.

17. Which of the following is the best summary of the simple effects analyses?
    a. Among the people who exercised, those who meditated were more optimistic than those who did not meditate. Among the people who did not exercise, those who meditated were not more optimistic than those who did not meditate.
    b. Among the people who exercised, those who meditated were more optimistic than those who did not meditate. Among the people who did not exercise, those who meditated were less optimistic than those who did not meditate.

Main Effect of Exercise

18. Graph the main effect of exercise using a bar graph. (Hint: You need to graph marginal means.)

19. Fill in the numbers for the results of the exercise main effect:

    $F(___, ___) = _____$, $p = _____$, $MSE = _____$, $\eta_p^2 = _____$.

20. Is the exercise main effect statistically significant?
    a. Yes
    b. No

21. Compute the effect size (d) comparing the exercise and no-exercise groups.

22. Select the best summary of the exercise main effect.
    a. The exercise group was more optimistic than the no-exercise group.
    b. The exercise group was less optimistic than the no-exercise group.
    c. The exercise and no-exercise groups did not differ in optimism.

Main Effect of Meditation

23. Graph the main effect of meditation using a bar graph.

24. Fill in the numbers for the results of the meditation main effect:

    $F(___, ___) = _____$, $p = _____$, $MSE = _____$, $\eta_p^2 = _____$.

25. Is the meditation main effect statistically significant?
    a. Yes
    b. No

26. Compute the effect size (d) comparing the meditation and no-meditation groups.

27. Select the best summary of the meditation main effect.
    a. The meditation group was more optimistic than the no-meditation group.
    b. The meditation group was less optimistic than the no-meditation group.
    c. The meditation and no-meditation groups did not differ in optimism.

28. Fill in the blanks for the APA-style summary of these data.

	Meditation		*No Meditation*		*Exercise Group Main Effect*
	n	***M* (*SD*)**	***n***	***M* (*SD*)**	***M* (*SD*)**
Exercise	10	6.90 (2.33)	10	3.50 (1.58)	5.20 (2.61)
No exercise	10	4.00 (1.41)	10	3.10 (2.03)	_____(_____)
Meditation group main effect		5.45 (2.40)		_____ (_____)	

A two-factor ANOVA with exercise group and meditation group as the grouping variables and optimism as the dependent variable revealed a significant interaction, $F(1, _____) = 4.45$, $p =$ _____, $MSE = 3.51$, $\eta_p^2 = .11$. Among the people who exercised, those who meditated were _____optimistic than those who did not meditate, $d =$ _____. Among the people who did not exercise, those who meditated were not more optimistic than those who did not meditate, $d = .52$.

The main effect of meditation group was also significant, $F(1, 36) =$ _____, $p < .001$, $MSE = 3.51$, $\eta_p^2 = .27$. Overall, people who meditated were more optimistic than those who did not meditate, $d = 1.03$.

Finally, the main effect of exercise group was _____, $F(1, 36) = 7.76$, $p = .001$, $MSE = 3.51$, $\eta_p^2 =$ _____. Overall, people who exercised were _____optimistic than those who did not exercise, $d =$ _____.

SCENARIO 2

The results of the previous study suggest that exercise and meditation are associated with optimism. However, it is not clear if exercise and meditation caused this greater optimism. It is entirely possible that people who exercise and meditate are just naturally more optimistic than those who do not engage in these activities. To determine if exercise and meditation lead to (or cause) a greater optimism, you need to randomly assign people to groups. Thus, in this next study, rather than find people who meditate or exercise, you find people who currently do neither and randomly assign them to one of the four groups: regular meditation/regular exercise, regular meditation/no exercise, no meditation/regular exercise, and no meditation/no exercise.

Suppose that you recruit 48 people for this study and randomly assign each person to one of the four groups. The output follows:

Descriptive Statistics

Dependent Variable: Optimism

MeditationGroup	ExerciseGroup	Mean	Std. Deviation	N
Meditation	Exercise	6.92	2.234	12
	No Exercise	3.67	2.060	12
	Total	5.29	2.678	24
No Meditation	Exercise	4.00	2.335	12
	No Exercise	2.58	1.975	12
	Total	3.29	2.236	24
Total	Exercise	5.46	2.686	24
	No Exercise	3.13	2.050	24
	Total	4.29	2.641	48

Tests of Between-Subjects Effects

Dependent Variable: Optimism

Source	Type III Sum of Squares	df	Mean Square	F	Sig	Partial Eta Squared
Corrected Model	123.417[a]	3	41.139	8.851	000	376
Intercept	884.083	1	884.083	190.218	000	812
MeditationGroup	48.000	1	48.000	10.328	002	190
ExerciseGroup	65.333	1	65.333	14.057	001	242
MeditationGroup * ExerciseGroup	10.083	1	10.083	2.170	148	047
Error	204.500	44	4.648			
Total	1212.000	48				
Corrected Total	327.917	47				

a. R Squared = .376 (Adjusted R Squared = .334)

29. Is the Meditation × Exercise interaction significant?
    a. Yes, the *p* value is equal to .47, which is less than .05.
    b. No, the *p* value is equal to .15, which is greater than .05.

30. Why isn't there a table of simple effects provided above?
    a. The interaction was not significant, so you know the simple effects are not significantly different from each other.
    b. I need to compute simple effects by hand.

31. Why do you need to compute effect sizes (*d*) for the simple effects even though the interaction was not significant?
    a. To determine if the main effects are significant
    b. To determine how large the effects are and if the study should be replicated with a larger sample

32. Compute the mean difference for the simple effect of exercise across the meditation and no-meditation conditions.

33. Compute the effect size (*d*) for the simple effect of exercise across the meditation and no-meditation conditions.

34. Compute the mean difference for the simple effect of no exercise across the meditation and no-meditation conditions.

35. Compute the effect size (*d*) for the simple effect of no-exercise group across the meditation and no-meditation conditions.

36. What do you think the next step should be for the researchers?
    a. Use a larger alpha to increase power.
    b. Redo the study with a larger sample size to increase power.

37. Is the main effect of exercise significant?
    a. Yes, the *p* value is equal to .001, which is less than .05.
    b. No, the *p* value is equal to .24, which is greater than .05.

38. Which group had higher optimism scores?
    a. The exercise group
    b. The no-exercise group

39. Is the main effect of meditation significant?
    a. Yes, the *p* value is equal to .002, which is less than .05.
    b. No, the *p* value is equal to .19, which is greater than .05.

40. Which group had higher optimism scores?
    a. The meditation group
    b. The no-meditation group

41. Fill in the missing values in the table and then write an APA style reporting statement below. Be sure to include the interaction and both main effects in your summary.

Table 1. Cell and Marginal Means and Standard Deviations

	Meditation		*No Meditation*		*Exercise Group Main Effect*
	n	***M (SD)***	***n***	***M (SD)***	***M (SD)***
Exercise	_____	_____	____	_____	_____
No exercise	_____	_____	_____	_____	_____
Meditation group main effect		_____		_____	

Activity 12.4: One-Way and Two-Way ANOVA Review

Learning Objectives

- Describe the difference between a one-way ANOVA and a two-way ANOVA
- Summarize ANOVA results from SPSS output files using APA style
- Describe an interaction effect

The research scenarios described below are based on Savani and Rattan (2012).

STUDY 1

Most American voters and economists think that dramatic income inequality is bad for a nation (Norton & Ariely, 2011). For example, most Americans disapprove of the fact that as of 2007, the richest 20% of people in the United States owned 85% of all the wealth in the country (Wolff, 2010). Two political analysts wanted to know if attitudes toward income inequality differed across Democrat, Independent, or Republican voters. They obtained a sample of 26 voters from each political party and asked them several questions like, "How disturbed are you by the fact that 20% of people in the United States own 85% of all the wealth in the country?" The voters responded by using a 7-point Likert scale with 1 indicating *not at all disturbed* and 7 indicating *extremely disturbed* for all of the questions. The researchers created an average "disturbed score" for each participant and ran a **one-way ANOVA.** The output follows.

Descriptive Statistics

Dependent Variable:disturbed_score

Party	Mean	Std. Deviation	N
Democrat	4.9789	.43790	26
Independent	4.2308	.48396	26
Republican	3.9846	.48554	26
Total	4.3981	.62924	78

Tests of Between-Subjects Effects

Dependent Variable:disturbed_score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	13.944[a]	2	6.972	31.609	.000	.457
Intercept	1508.776	1	1508.776	6840.222	.000	.989
Party	13.944	2	6.972	31.609	.000	.457
Error	16.543	75	.221			
Total	1539.264	78				
Corrected Total	30.487	77				

a. R Squared = .457 (Adjusted R Squared = .443)

Estimates

Dependent Variable:disturbed_score

Party	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Democrat	4.979	.092	4.795	5.162
Independent	4.231	.092	4.047	4.414
Republican	3.985	.092	3.801	4.168

Multiple Comparisons

disturbed_score

Tukey HSD

(I) Party	(J) Party	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Democrat	Independent	.7481*	.13026	.000	.4367	1.0596
	Republican	.9943*	.13026	.000	.6828	1.3058
Independent	Democrat	-.7481*	.13026	.000	-1.0596	-.4367
	Republican	.2462	.13026	.149	-.0653	.5576
Republican	Democrat	-.9943*	.13026	.000	-1.3058	-.6828
	Independent	-.2462	.13026	.149	-.5576	.0653

Based on observed means.

The error term is Mean Square(Error) = .221.

*. The mean difference is significant at the .05 level.

1. How many independent variables (IVs) are in the preceding study?
   a. 1
   b. 2
   c. 3

2. Identify the IV(s) in this study.
   a. Political party
   b. Republican
   c. Democrat
   d. Independent
   e. Rating on scale measuring how disturbed they are by income inequality
   f. Income inequality in the United States

3. What scale of measurement is this (are these) independent variable(s) measured on (i.e., nominal, ordinal, interval, or ratio)?
   a. Nominal
   b. Ordinal
   c. Interval/ratio

4. Identify the DV in this study.
   a. Political party
   b. Republican
   c. Democrat
   d. Independent
   e. Rating on scale measuring how disturbed people are by income inequality
   f. Income inequality in the United States

5. What scale of measurement is the dependent variable measured on (i.e., nominal, ordinal, interval, or ratio)?
   a. Nominal
   b. Ordinal
   c. Interval/ratio

6. After looking at the above SPSS output, the two political analysts disagree on how to interpret the results. Bob says that there is a statistically significant interaction, but Alisha says that there can't be an interaction. Which researcher is correct, and why?
   a. Bob is correct because the *p* value is less than .05 and the post hoc tests indicate that some of the pairwise comparisons are significant.
   b. Alisha is correct because you can only have an interaction if there are two or more independent variables and this study has just one IV.

7. Based on the above output, is there a statistically significant difference between those in the different political parties on the degree to which they are disturbed by income inequality in the United States?
   a. Yes, the *p* value for the mean difference across political party was less than .05.
   b. No, the effect size for political party was greater than .05.

8. Which pairwise comparisons are statistically significant? (Choose all that apply.)
   a. Republicans and Democrats
   b. Republicans and Independents
   c. Democrats and Independents

9. Use the above SPSS output to compute the effect sizes (*d*s) for each of the pairwise comparisons.

   d for Republican vs. Democrat difference =

   d for Republican vs. Independent difference =

   d for Democrat vs. Independent difference =

10. Which group was *most* disturbed by income inequality?
    a. Republicans
    b. Democrats
    c. Independents

11. A brief report of the results in APA format is below, but some of the necessary statistical information is missing. Use the provided SPSS output to fill in the necessary information. Furthermore, there are two errors in the write up. Fix these errors.

Party	*Mean (Standard Deviation)*
Republican	3.98 (.49)
Democrat	4.98 (.44)
Independent	4.23 (.48)

A two-factor ANOVA was conducted with political party as the independent variable and perceptions of income inequality as the dependent variable. Overall, there was a significant interaction, indicating that those from different political parties differed in the degree to which they were disturbed by the income inequality in the United States, F(___, ____) = ___, p _____, MSE = _____, η_p^2 = ____. Two of the three political party comparisons were significantly different. Democrats were more disturbed by the income inequality than Independents, p _____, d = _____. They were also more disturbed by the income inequality than Republicans, p _____, d = _____. Independents and Republicans were not significantly different, p = _____, d = _____, although the medium effect size suggests that the study should be replicated with a larger sample size.

STUDY 2

Personal choice is a central concept of American culture. The idea that individuals should, at least to some degree, succeed or fail based on the quality of their personal choices is a belief commonly held by many Americans. The political analysts want to know if they could make Americans more tolerant of income inequality in the United States by highlighting Americans' strongly held belief in "personal choice." To answer this question, the operatives obtained a sample of 26 Democrats, 26 Independents, and 26 Republicans. All the voters watched the same 6-minute video, but half of the voters from each party were instructed to note each time the actor in the video made a "choice," while the other half of the voters in each party noted every time the actor touched something. After watching the video, all voters provided responses to questions like "How disturbed are you by the fact that, in 2007, the richest 20% of people in the United States own 85% of all wealth in the country." The voters responded by using a 7-point Likert scale with 1 indicating *not at all disturbed* and 7 indicating *extremely disturbed*. The political analysts reasoned that the "personal choice" group would be less disturbed by income inequality because they would be more likely to conclude that the richest people made better personal choices than the rest of the population and therefore that they deserved their wealth. The analysts ran a **two-way ANOVA**, and the SPSS output follows.

Descriptive Statistics

Dependent Variable:disturbed_score2

Party	Instructions	Mean	Std. Deviation	N
Democrat	choice	4.4846	.49807	13
	touch	4.9568	.47262	13
	Total	4.7207	.53317	26
Independent	choice	3.7846	.58715	13
	touch	4.2308	.38381	13
	Total	4.0077	.53660	26
Republican	choice	3.4769	.37451	13
	touch	4.1077	.59507	13
	Total	3.7923	.58373	26
Total	choice	3.9154	.64340	39
	touch	4.4318	.61069	39
	Total	4.1736	.67518	78

Tests of Between-Subjects Effects

Dependent Variable:disturbed_score2

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	17.608[a]	5	3.522	14.493	.000	.502
Intercept	1358.659	1	1358.659	5591.729	.000	.987
Party	12.278	2	6.139	25.267	.000	.412
Instructions	5.200	1	5.200	21.399	.000	.229
Party * Instructions	.130	2	.065	.267	.766	.007
Error	17.494	72	.243			
Total	1393.761	78				
Corrected Total	35.102	77				

a. R Squared = .502 (Adjusted R Squared = .467)

Estimates

Dependent Variable:disturbed_score2

Party	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Democrat	4.721	.097	4.528	4.913
Independent	4.008	.097	3.815	4.200
Republican	3.792	.097	3.600	3.985

Pairwise Comparisons

Dependent Variable:disturbed_score2

(I) Party	(J) Party	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
Democrat	Independent	.713*	.137	.000	.440	.986
	Republican	.928*	.137	.000	.656	1.201
Independent	Democrat	-.713*	.137	.000	-.986	-.440
	Republican	.215	.137	.120	-.057	.488
Republican	Democrat	-.928*	.137	.000	-1.201	-.656
	Independent	-.215	.137	.120	-.488	.057

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

12. How many independent variables (IVs) are in the above study?
    a. 1
    b. 2
    c. 3

13. Identify the IV(s) in this study.
    a. Political party
    b. Republican
    c. Democrat
    d. Independent
    e. Action performed while watching video

f. Notice number of personal choices
g. Notice number of touches
h. Rating on scale measuring how disturbed they are by income inequality
i. Income inequality in the United States

14. What scale of measurement is this (are these) independent variable(s) measured on (i.e., nominal, ordinal, interval, or ratio)?
    a. Nominal
    b. Ordinal
    c. Interval/ratio

15. Identify the DV in this study.
    a. Political party
    b. Republican
    c. Democrat
    d. Independent
    e. Rating on scale measuring how disturbed they are by income inequality
    f. Income inequality in the United States

16. What scale of measurement is this dependent variable measured on (i.e., nominal, ordinal, interval, or ratio)?
    a. Nominal
    b. Ordinal
    c. Interval/ratio

17. Why can a two-way ANOVA produce a significant interaction between two IVs while a one-way ANOVA can never produce an interaction?
    a. An interaction only occurs when one IV interacts with a DV. One-way ANOVAs have both an IV and a DV, so interactions are possible.
    b. An interaction only occurs when two IVs interact with the DV. Two-way ANOVAs have two IVs, so they can result in an interaction while one-way ANOVAs have just one IV and cannot result in an interaction.
    c. An interaction only occurs when two IVs combine to have a unique effect on the DV. One-way ANOVAs only have one IV, so they cannot have an interaction, while two-way ANOVAs have two IVs.

18. Does this study have a significant interaction?
    a. Yes, the p value for the interaction is .007, which is less than .05.
    b. No, the p value for the interaction is .766, which is greater than .05.
    c. Yes, some of the pairwise comparisons are statistically significant.
    d. No, the F is in the critical region.

19. Graph the interaction below.

20. What about the graph indicates that there is *no* interaction between the IVs? Select all that apply.
    a. The effect of instructions while watching the video is the same, regardless of political party.
    b. The choice condition resulted in lower ratings than the touch condition for Republicans, Democrats, and Independents.
    c. The choice condition had the largest effect on Independents, followed by Democrats and then Republicans.

21. As you know, a two-way ANOVA produces three *F* ratios. Find the statistical information for the interaction effect and fill in the following blanks.

    F(____, ____) = _____, p = _____, MSE = _____, η_p^2 = ____.

22. The other two *F* ratios will always test the main effect of the first IV and main effect of the second IV, respectively. In the spaces provided below, list the two IVs, the specific *F* ratio that tests the main effect of each IV, and the *p* value for each main effect.

    IV 1 = _____;

    F(____, ____) = _____, p = _____, MSE = _____, η_p^2 = ____.

    IV 2 = _____;

    F(____, ____) = _____, p = _____, MSE = _____, η_p^2 = ____.

23. Record the mean and standard deviation for each political party.
    a. Republican: M = _____, SD = _____
    b. Democrat: M = _____, SD = _____
    c. Independent: M = _____, SD = _____

24. Based on the above output, is there a statistically significant main effect of political party? In other words, is there a statistically significant difference between those in the different political parties in the degree to which they are disturbed by income inequality in the United States?
    a. No, the *p* value for the main effect of political party is greater than .05.
    b. Yes, the *p* value for the main effect of political party is less than .05.

25. Based on the above output, which political parties were significantly different in terms of their "disturbed by income inequality scores"? (Choose all that apply)
    a. Republicans and Democrats
    b. Republicans and Independents
    c. Democrats and Independents

26. Compute the effect sizes (*d*s) for the each of the pairwise political party differences.

    d for Republicans vs. Democrats difference =

d for Republicans vs. Independents difference =

d for Democrats vs. Independents difference =

27. Record the mean and standard deviation for the touch and choice groups.

    Choice group: $M =$ _____, $SD =$ _____

    Touch group: $M =$ _____, $SD =$ _____

28. Based on the above output, is there a statistically significant main effect of instructions? In other words, is there a significant difference between those who noted "personal choice" while watching the video and those who noted "touches" while watching the video on the degree to which they are disturbed by income inequality in the United States?
    a. No, the *p* value for the main effect of instructions is greater than .05.
    b. Yes, the *p* value for the main effect of instructions is less than .05.

29. Why don't you need to look at post hoc tests for the main effect of instructions?
    a. Because the main effect was not significant
    b. Because there are only two groups and you can just look to see which mean is higher

30. Compute the effect size (*d*) for main effect of instructions.

d for personal choice vs. touches difference =

31. The following APA-style report includes three errors or omissions (one in each paragraph). Locate and fix these errors. Also, fill in all missing numbers.

    A two-factor ANOVA was conducted with political party and instructions as the independent variables and perceptions of income inequality as the dependent variable. The interaction

Table 1. Cell and Marginal Means and Standard Deviations

	Choice		*Touch*		*Political Party Main Effect*
	n	***M (SD)***	***n***	***M (SD)***	***M (SD)***
Democrat	_____	_____	_____	_____	_____
Republican	_____	_____	_____	_____	_____
Independent	_____	_____	_____	_____	_____
Choice main effect	_____		_____		

between political party and instructions was not significant, $F(2, _____) = .27$, $p = .76$, $MSE = .24$, $\eta_p^2 = .01$. The instructions had different effects on every political party.

There was a significant main effect for political party, $F(2, 72) = 25.27$, $p < .001$, $MSE = .24$, $\eta_p^2 =$ _____. Overall, Democrats were more disturbed by income inequality than were Republicans, $p =$ _____, $d =$ _____ and Independents, $p =$ _____, $d =$ _____.

Finally, there was a significant main effect for instructions, $F(2, 72) = 21.40$, $p < .001$, $MSE =$_____, $\eta_p^2 = .23$. Overall, those who were attending to choices were significantly different from those who attended to "touches," $p =$ _____, $d =$ _____.

STUDY 3

Now the political analysts want to use the information they learned about highlighting personal choice to create an effective political advertisement. They want to create an advertisement that will make a proposed tax cut for those making over $10 million a year more palatable to middle-class voters. In the advertisement, a well-known politician talks about all the positive personal choices successful businesspeople make every day and how these choices have positive results for their local communities. The analysts get a sample of 40 Independents and 40 Democrats and show the political ad to half of the people from each political party. Then all of the voters answer questions indicating how likely they are to support the proposed tax cut using a 7-point Likert scale with 1 indicating *not at all likely* and 7 indicating *extremely likely*. The operatives create an "average support score" and run a **two-way ANOVA** and the SPSS output follows.

Descriptive Statistics

Dependent Variable: Tax_support_score

Party	Video_Condition	Mean	Std. Deviation	N
Democrat	Watched Ad	2.2524	.50704	20
	Did not watch Ad	2.4984	.73467	20
	Total	2.3754	.63539	40
Independent	Watched Ad	2.9290	.48704	20
	Did not watch Ad	2.5650	.43653	20
	Total	2.7470	.49231	40
Total	Watched Ad	2.5907	.59849	40
	Did not watch Ad	2.5317	.59743	40
	Total	2.5612	.59490	80

Tests of Between-Subjects Effects

Dependent Variable: Tax_support_score

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	4.691[a]	3	1.564	5.108	.003	.168
Intercept	524.774	1	524.774	1714.120	.000	.958
Party	2.761	1	2.761	9.019	.004	.106
Video_Condition	.070	1	.070	.227	.635	.003
Party * Video_Condition	1.861	1	1.861	6.078	.016	.074
Error	23.267	76	.306			
Total	552.733	80				
Corrected Total	27.959	79				

a. R Squared = .168 (Adjusted R Squared = .135)

Pairwise Comparisons

Dependent Variable: Tax_support_score

Party	(I) Video_Condition	(J) Video_Condition	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
						Lower Bound	Upper Bound
Democrat	Watched Ad	Did not watch Ad	-.246	.175	.164	-.595	.102
	Did not watch Ad	Watched Ad	.246	.175	.164	-.102	.595
Independent	Watched Ad	Did not watch Ad	.364*	.175	.041	.016	.712
	Did not watch Ad	Watched Ad	-.364*	.175	.041	-.712	-.016

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

32. This two-way ANOVA generated three F ratios, each one testing a specific effect. In the space provided below, list the specific effects this analysis tested, the value of the F ratio testing each effect, and the p value associated with each F ratio.

Main effect of _____;

$F(____, ____) = _____, p = _____, MSE = _____, \eta_p^2 = ____.$

Main effect of _____;

$F(____, ____) = _____, p = _____, MSE = _____, \eta_p^2 = ____.$

Interaction between_____;

$F(____, ____) = _____, p = _____, MSE = _____, \eta_p^2 = ____.$

33. What does it mean to say that an effect is "significant"? Saying that an effect is significant means (choose all that apply)

    a. that it is important; all significant results are important and only significant results can be important.

    b. that the mean difference(s) compared by the effect are so large that they were unlikely to have occurred by chance.

    c. that the researcher should redo the study with a larger sample size.

    d. that the effect size for that effect is moderate or large.

34. Indicate which of the following effects were "significant." (Choose all that apply.)

    a. The main effect of political party

    b. The main effect of video condition

    c. The interaction between political party and video condition

35. After identifying which effects were and were not "significant," you should identify the meaning of each effect by visually inspecting the means compared by each effect. What is the meaning of the main effect of party in this study?

    a. Independents and Democrats were equally favorable toward the proposed tax cut.

    b. Independents were more favorable toward the proposed tax cut than were Democrats.

    c. Democrats were more favorable toward the proposed tax cut than were Independents.

36. What is the meaning of the main effect of video in this study?

    a. Overall, the video did not make voters more favorable toward the proposed tax cut.

    b. The video did make voters more favorable toward the proposed tax cut.

    c. The video made voters less favorable toward the proposed tax cut.

37. What is the meaning of the interaction effect in this study?

    a. The video affected Independents and Democrats the same way, meaning that there was no interaction.

    b. The video increased Democrats' attitudes toward the proposed tax cuts more than it increased Independents' scores.

    c. The video increased Independents' attitudes toward the proposed tax cuts more than it increased Democrats' scores.

38. The following APA-style report includes six errors or omissions (two in each paragraph). Locate and fix these errors. Also, fill in all missing numbers

    A two-factor ANOVA was conducted with political party and video condition as the independent variables and attitudes toward tax cuts as the dependent variable. There was a significant interaction between political party and video condition, $F(1, 76) = 6.08$, $p = .02$, $MSE = .31$, $\eta_p^2 = .07$. For Democrats, those who watched the video and those who did not had _____ scores, suggesting that the video had little to no effect on attitude toward the proposed tax cut, $p =$ _____, $d =$ _____. However, the video was less effective on Independents. Independents who watched

Table 1. Cell and Marginal Means and Standard Deviations

	Watched Ad		*Did Not Watch Ad*		*Political Party Main Effect*
	n	***M (SD)***	***n***	***M (SD)***	***M (SD)***
Democrat	20	_____	_____	2.50 (.73)	2.38 (.64)
Independent	20	_____	_____	_____	_____
Video main effect		2.59 (.60)		_____	

the video were more favorable toward the proposed tax cut than Independents who did not watch the video, $p =$ _____, $d =$ _____.

There was a significant interaction for political party, $F(1, 76) = 9.02$, $p = .004$, $MSE = .31$, $\eta_p^2 = .11$. Overall, Democrats were more supportive of the proposed tax cuts than Independents, $d =$ _____.

Finally, there was a significant main effect for video condition, $F(1, 72) = .23$ $p = .64$, $MSE = .31$, $\eta_p^2 = .003$. Overall, those who watched the video were more supportive of tax cuts than those who did not watch the video, $d =$ _____.

Activity 12.5: Choose the Correct Statistic

Learning Objectives

After reading the chapter and completing the homework and this activity, you should be able to do the following:

- Read a research scenario and determine which statistic should be used
- The two-factor ANOVA is the last statistic in this book that can be used to compare means. Any research scenario that involves two separate independent variables and one dependent variable requires a two-way ANOVA. You should also note that any time you want to compare three or more means, you will need to use an ANOVA. If there is just one IV, you use a one-way ANOVA, and if there are two IVs, you use a two-way ANOVA. Use the table and flowchart in Appendix J to help you determine which statistics should be used to answer the research question.

CHOOSE THE CORRECT TEST STATISTIC

Determine which of the following statistics should be used in each of the following research scenarios: *z* for sample mean, single-sample *t*, related samples *t*, independent samples *t*, one-way independent samples ANOVA, or two-way independent samples ANOVA.

1. Mark Haub, a professor at Kansas State University, reported that he lost 27 pounds in 2 months by consuming 1,800 calories a day of primarily junk food (e.g., Twinkies). Given the low caloric intake, the weight loss was not surprising, but he reported that his cholesterol levels also improved. A nutritionist wonders what the long-term effects of a junk food diet are and so designs an experiment using rats. The cholesterol levels of a group of 50 rats are measured, and then the rats are fed a diet of Twinkies and Doritos for 1 year. At the end of that year, the rats' cholesterol levels are measured again. Which statistic should be used to determine if the diet had a significant effect on cholesterol levels?

2. Another researcher reads of Professor Haub's experience and learns that he consumed one serving of vegetables a day while on this diet and wonders if this small amount of vegetables was an important factor in lowering his cholesterol levels. To test the effects of this diet plus vegetables on cholesterol levels, she obtains a random sample of 75 people and randomly assigns them to eat 1,800 calories a day of only junk food, 1,800 calories a day of junk food and one serving of vegetables, or 1,800 calories a day of only vegetables. Cholesterol levels were measured after 2 months on the diet. Which statistic should be used to determine if there is a significant difference in cholesterol levels among the three different diets?

3. A number of web reports indicate that Professor Haub's experience suggests that the quality of food is not as important as the amount of food people consume. A nutritionist decides to test this by randomly assigning 100 people with 25% body fat to eat either 1,800 calories a day or 2,500 calories a day. He also manipulates the quality of the food. Half of the participants eat only junk food, and the other half eat a healthy diet consisting of fruits, vegetables, whole grains, and lean protein. After 2 months on this diet, the percentage of body fat was recorded for each participant. Which statistic should the nutritionist use to determine that the effect of caloric intake on body fat percentage is different for junk food and for healthy food diets?

4. A statistics teacher thinks that doing homework improves scores on statistics exams. To test this hypothesis, she randomly assigns students to two groups. One group is required to work on the homework until all problems are correct. Homework is optional for the second group. At the end of the semester, final grades are compared between the two groups, and the results reveal that the required-homework group had higher final grades than the optional-homework group. Which statistic should be used to compare the final grades of those in each homework group?

5. Encouraged by the results of the first study, the statistics teacher wonders if it is necessary to complete the entire homework assignment until it is correct. Perhaps just working on the homework is sufficient to improve grades. Thus, the following semester, she randomly assigns students to three groups: (1) homework is optional, (2) must get all homework questions correct, and (3) need to answer each homework question but they do not have to be correct. At the end of

the semester, final grades are compared between the three groups. What statistic could be used to compare the final grades of the three groups?

6. A recent study revealed that students from the University of Washington who studied abroad reported that they drank more alcohol while studying abroad than they used to before they studied abroad. Which statistic should be used to make this comparison?

7. A college counselor does a survey of students and finds that students on her campus are getting an average of just 6.5 hours of sleep on weekdays. In an attempt to get the students to sleep more, the counseling center sends an e-mail to all students on campus. This e-mail discusses the benefits of sleep as well as the dangers of sleep deprivation. In addition to the e-mail, the counseling center gives presentations about the value of sleep to all incoming freshmen. Later in the semester, a random sample of 78 students is asked to report the number of hours of sleep they get on an average weekday. They report an average of 7.1 hours of sleep, with a standard deviation of 0.9. Did the intervention increase the mean amount of sleep on campus?

8. Twenty-six students were identified as needing additional help with reading comprehension. These students took an extra class during summer that specifically focused on understanding what they were reading in short stories, novels, and newspaper articles. After the summer course, the students' average score was 70. What statistic should be used to determine if the mean for this sample of students is significantly different from the mean for the population ($\mu = 75$ and $\sigma = 10$)?

9. A person with schizophrenia is more likely to suffer a relapse if family members are highly critical, hostile, and overinvolved in that person's life. Some psychologists wondered if training the family members of schizophrenic patients to be less critical and less hostile would reduce the patients' symptoms. The psychologists divided a sample of people with schizophrenia into two groups. The family members of the first group were trained to reduce their critical and hostile interaction patterns. The family members of the second group of patients received no training. Six months later, patients in both groups were evaluated and given scores that reflected the severity of their schizophrenic symptoms. Which statistic should be used to determine if there is a significant difference between the severity of the schizophrenic symptoms across the trained and no-training groups?

10. In a related study, researchers divided a sample of family members of schizophrenic patients into two groups. One group of family members was trained to be less critical. The other group of family members was not trained at all. In addition, half of the people in each group of family members were male, and half were female. Is the training program equally effective at reducing critical verbal comments made by males and by females?

11. Based on a previous research study, it is known that the mean number of critical comments made by family members of schizophrenic patients in a day is $M = 28.2$. A researcher wanted to know if family members of schizophrenic patients make more critical comments than people who do not have a family member who is diagnosed with schizophrenia. The researcher obtained a sample of people who do not have a family member diagnosed with schizophrenia and found that the number of critical comments made in a day by these people was $M = 25.3$, with a standard deviation of 2.5. Which statistic should be used to determine if the number of critical comments was significantly different in schizophrenic and nonschizophrenic families?

Chapter 12 Practice Test

A researcher decides to investigate if the best type of insect repellant depends on one's personal body temperature. Specifically, she investigates the relative effectiveness of an insect repellant containing DEET versus an insect repellant containing natural ingredients (e.g., lemon eucalyptus, and citronella) for people who have high versus low body temperatures. The researcher uses procedures similar to those used by previous studies investigating mosquito bites. The following data are the number of mosquito bites suffered by those in each experimental condition.

Low Body Temperature		*High Body Temperature*	
Natural	*DEET*	*Natural*	*DEET*
5	6	12	10
6	5	11	9
7	7	13	11
8	8	14	12
6	7	15	13
4	6	17	15
7	7	18	11

1. Match the assumption to the fact that is relevant to that assumption.

   _____ Independence

   _____ Appropriate measurement of the IV and the DV

   _____ Normality

   _____ Homogeneity of variance

   a. Data were collected from one participant at a time.
   b. The one variable defines groups and the participants' responses were given on an interval/ratio scale.
   c. Samples of 30 or more tend to form distributions of sample means that meet this assumption; also, if the population of scores has a normal shape, this assumption will be met.
   d. When the standard deviations from each condition are similar (i.e., not twice as large), this assumption is probably met.

2. If you entered the above data into an SPSS file the correct way to run a two-way ANOVA, how many *columns* would your data file have?

   a. 1
   b. 2

c. 3

d. 4

3. If you entered the above data into an SPSS file the correct way to run a two-way ANOVA, how many *rows* would your data file have?

a. 7

b. 14

c. 21

d. 28

The SPSS output for these data follow.

Descriptive Statistics

Dependent Variable: Bites

BodyTemp	Repellant	Mean	Std. Deviation	N
low temp	natural	6.1429	1.34519	7
	DEET	6.5714	.97590	7
	Total	6.3571	1.15073	14
high temp	natural	14.2857	2.56348	7
	DEET	11.5714	1.98806	7
	Total	12.9286	2.61547	14
Total	natural	10.2143	4.66045	14
	DEET	9.0714	2.99908	14
	Total	9.6429	3.88934	28

Tests of Between-Subjects Effects

Dependent Variable: Bites

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	328.714[a]	3	109.571	32.989	.000	.805
Intercept	2603.571	1	2603.571	783.871	.000	.970
BodyTemp	302.286	1	302.286	91.011	.000	.791
Repellant	9.143	1	9.143	2.753	.110	.103
BodyTemp * Repellant	17.286	1	17.286	5.204	.032	.178
Error	79.714	24	3.321			
Total	3012.000	28				
Corrected Total	408.429	27				

a. R Squared = .805 (Adjusted R Squared = .780)

Pairwise Comparisons

Dependent Variable: Bites

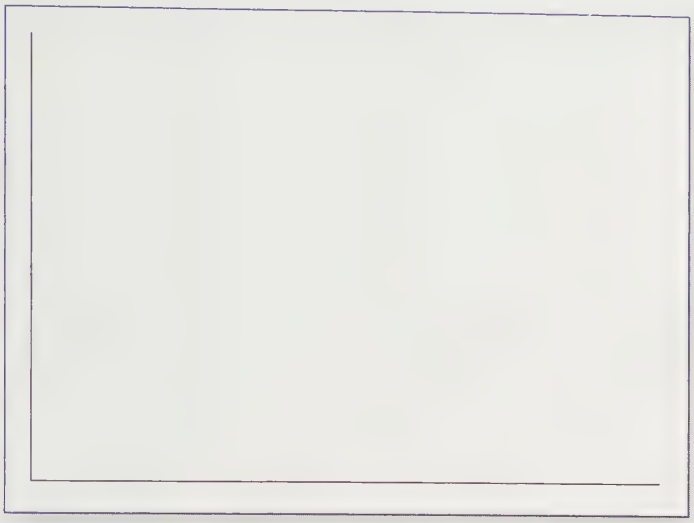
BodyTemp	(I) Repellant	(J) Repellant	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
						Lower Bound	Upper Bound
low temp	natural	DEET	-.429	.974	.664	-2.439	1.582
	DEET	natural	.429	.974	.664	-1.582	2.439
high temp	natural	DEET	2.714*	.974	.010	.704	4.725
	DEET	natural	-2.714*	.974	.010	-4.725	-.704

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

4. Which of the following is the best verbal summary of the null hypothesis for the interaction?
   a. The effect of type of insect repellant is the same for people with high and low body temperatures.
   b. The effect of type of insect repellant is different for people with high and low body temperatures.
   c. Those with higher and lower body temperatures will have similar numbers of mosquito bites.
   d. The mean number of bites for the two types of insect repellant will be similar.
5. Which of the following is the best verbal summary of the research hypothesis for the interaction?
   a. The effect of type of insect repellant is the same for people with high and low body temperatures.
   b. The effect of type of insect repellant is different for people with high and low body temperatures.
   c. Those with higher and lower body temperatures will have different numbers of mosquito bites.
   d. The mean number of bites for the two types of insect repellant will be different.
6. The interaction tests the difference between
   a. the marginal means.
   b. the two main effects.
   c. specific pairs of cell means.
7. Graph the interaction between body temperature and repellant type.



8. Report the F statistic information for the interaction.

   F(_____, _____) = _____, p = _____, MSE = _____, η_p^2 = _____.

9. You should reject the null hypothesis if (choose all that apply)
   a. the p value is less than the alpha.
   b. the p value is greater than the alpha.
   c. the obtained F statistic is greater than the critical value.
   d. the obtained F statistic is less than the critical value.

10. Should you reject the null hypothesis for the interaction?
    a. Yes, reject the null for the interaction.
    b. No, fail to reject the null for the interaction.

11. Based on the simple effects analysis in the SPSS output, which of the following statements are true? Select all that apply.
    a. For low body temperature participants, there was no meaningful difference between repellants containing DEET and natural ingredients.
    b. For low body temperature participants, repellants containing natural ingredients were slightly better than those containing DEET.
    c. For high body temperature participants, there was no meaningful difference between repellants containing DEET and natural ingredients.
    d. For high body temperature participants, repellants containing DEET were better than those containing natural ingredients.

12. What is the effect size, *d*, for the simple effect of *low* body temperature participants across natural ingredients versus DEET repellants?
    a. −.36
    b. −42
    c. −.68
    d. 1.18
    e. 2.71
    f. .98

13. What is the effect size, *d*, for the simple effect of *high* body temperature participants across natural ingredients versus DEET repellants?
    a. −.36
    b. −.42
    c. −.68
    d. 1.18
    e. 2.71
    f. .98

14. Which of the following best summarizes the results of the interaction?
    a. People with a higher body temperature ($M = 12.93$, $SD = 2.62$) were more likely to be bitten than people with a lower body temperature ($M = 6.36$, $SD = 1.15$), and people were more likely to be bitten when they used the natural insect repellant ($M = 10.21$, $SD = 4.66$) than when they used the repellant containing DEET ($M = 9.07$, $SD = 3.00$), $F(1, 24) = 5.20$, $p = .03$, $MSE = 3.32$, $\eta_p^2 = .18$.
    b. People with a higher body temperature were much less likely to be bitten when they used the insect repellant containing DEET ($M = 11.57$, $SD = 1.99$) than the natural insect repellant ($M = 14.29$, $SD = 2.56$), but for those with lower body temperatures, there was no significant difference in bites between those using the natural insect repellant ($M = 6.14$, $SD = 1.35$) and the repellant containing DEET ($M = 6.57$, $SD = .98$), $F(1, 24) = 5.20$, $p = .03$, $MSE = 3.32$, $\eta_p^2 = .18$.
    c. The interaction between body temperature and type of insect repellant was not statistically significant, $F(1, 24) = 5.20$, $p = .03$, $MSE = 3.32$, $\eta_p^2 = .18$.

15. Which of the following is the best verbal summary of the null hypothesis for the main effect of body temperature?
    a. The mean number of mosquito bites for the higher body temperature group will not be different from the mean number of mosquito bites for the lower body temperature group.
    b. The mean number of mosquito bites for the higher body temperature group will be different from the mean number of mosquito bites for the lower body temperature group.

16. Which of the following is the best verbal summary of the research hypothesis for the main effect of body temperature?
    a. The mean number of mosquito bites for the higher body temperature group will not be different from the mean number of mosquito bites for the lower body temperature group.
    b. The mean number of mosquito bites for the higher body temperature group will be different from the mean number of mosquito bites for the lower body temperature group.

17. Graph the main effect of body temperature.

18. Report the *F* statistic information for the main effect of body temperature.

F(_____, _____) = _____, p = _____, MSE = _____, η_p^2 = _____.

19. Should you reject the null hypothesis for the main effect of body temperature?
    a. Yes
    b. No

20. What is the effect size, *d*, for the main effect of body temperature?
    a. .98
    b. 6.57
    c. 3.25
    d. 1.54

21. Which of the following best summarizes the results for the main effect of body temperature?
    a. People with a higher body temperature (M = 12.93, SD = 2.62) were more likely to be bitten than were people with a lower body temperature (M = 6.36, SD = 1.15), $F(1, 24) = 91.01$, $p < .001$, $MSE = 3.32$, $\eta_p^2 = .79$.
    b. People with a higher body temperature (M = 12.93, SD = 2.62) were more likely to be bitten than were people with a lower body temperature (M = 6.36, SD = 1.15), and people were more likely to be bitten when they used the natural insect repellant (M = 10.21, SD = 4.66) than when they used the repellant containing DEET (M = 9.07, SD = 3.00), $F(1, 24) = 91.01$, $p < .001$, $MSE = 3.32$, $\eta_p^2 = .79$.
    c. The main effect of body temperature was not statistically significant, $F(1, 24) = 91.01$, $p < .001$, $MSE = 3.32$, $\eta_p^2 = .79$.

22. Which of the following is the best verbal summary of the null hypothesis for the main effect of type of insect repellant?
    a. The mean number of mosquito bites for people using the insect repellant containing DEET will not be different from the mean number of mosquito bites for people using the natural insect repellant.
    b. The mean number of mosquito bites for people using the insect repellant containing DEET will be different from the mean number of mosquito bites for people using the natural insect repellant.

23. Which of the following is the best verbal summary of the research hypothesis for the main effect of type of insect repellant?
    a. The mean number of mosquito bites for people using the insect repellant containing DEET will not be different from the mean number of mosquito bites for people using the natural insect repellant.
    b. The mean number of mosquito bites for people using the insect repellant containing DEET will be different from the mean number of mosquito bites for people using the natural insect repellant.

24. Graph the main effect of insect repellant type.

25. Report the *F* statistic information for the main effect of insect repellant type.

$F(____, ____) = ____, p = ____, MSE = ____, \eta_p^2 = ____.$

26. Should you reject the null hypothesis for the main effect of type of insect repellant type?
    a. Yes
    b. No

27. What is the effect size, *d*, for the main effect of insect repellant type?
    a. .29
    b. .15
    c. 1.14
    d. .82

28. Which of the following best summarizes the results for the main effect of type of insect repellant type?
    a. People were more likely to be bitten when they used the natural insect repellant ($M = 10.21$, $SD = 4.66$) than when they used the insect repellant containing DEET($M = 9.07$, $SD = 3.00$), $F(1, 24) = 2.75$, $p = .11$, $MSE = 3.32$, $\eta_p^2 = .10$.
    b. The main effect of type of insect repellant was not statistically significant, $F(1, 24) = 2.75$, $p = .11$, $MSE = 3.32$, $\eta_p^2 = .10$.

REFERENCES

Josefsson, T., Lindwall, M., & Archer, T. (2014). Physical exercise intervention in depressive disorders: Meta-analysis and systematic review. *Scandinavian Journal of Medicine & Science in Sports, 24*(2), 259–272.

Langer, P., Holzner, B., Magnet, W., & Kopp, M. (2005). Hands-free mobile phone conversation impairs the peripheral visual system to an extent comparable to an alcohol level of 4–5g 100 ml. *Human Psychopharmacology: Clinical and Experimental, 20*(1), 65–66.

Norton, M. I., & Ariely, D. (2011). Building a better America—One wealth quintile at a time. *Perspectives on Psychological Science, 6*(1), 9-12.

Roediger, H. I., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255.

Savani, K., & Rattan, A. (2012). A choice mind-set increases the acceptance and maintenance of wealth inequality. *Psychological Science, 23*(7), 796–804.

Wolff, E. N. (2010). Recent trends in household wealth in the United States: Rising debt and the middle-class squeeze—An update to 2007. *The Levy Economics Institute Working Paper Collection, 589*, 1–59.

CHAPTER 13

Correlation and Regression

LEARNING OBJECTIVES

After reading this chapter, you should be able to do the following:

- Identify when to use Pearson's and Spearman's correlations
- Interpret the sign and value of a correlation coefficient
- Draw and interpret scatterplots by hand and using SPSS
- Write null and research hypotheses using words and symbols
- Compute the degrees of freedom (*df*) and determine the critical region
- Compute a Pearson's and Spearman's correlation coefficients by hand (using a calculator)
- Determine whether or not you should reject the null hypothesis
- Compute an effect size (r^2) and describe it
- Summarize the results of the analysis using American Psychological Association (APA) style
- Interpret the SPSS correlation output

WHEN TO USE CORRELATIONS AND WHAT THEY CAN TELL YOU

All the test statistics you have learned so far compare means. For example, when doing an independent *t* test, you have one grouping variable (e.g., male vs. female) and one variable measured on an interval/ratio scale (e.g., test scores). A mean is computed for each group (i.e., males and females), and a statistical test is done to determine if the means are significantly different. Comparing means of different groups is a very common research approach. Another type of research that is very common tries to determine if two interval/ratio variables are related to each other. For example, is college grade point average (GPA) related to annual income? To answer this question, you could record the college GPA and annual income of a sample of people and then compute a statistic to determine if these two variables are related. Note that in this research approach, you are not comparing groups of people; instead, you use a **correlation coefficient** to determine if a high score on one variable (e.g., college GPA) is associated with a high or low score on another variable (e.g., annual income). Interpreting a correlation coefficient comes down to reading its "direction" and its "strength." A coefficient's *direction* is revealed by how individual participants' scores on the two variables tend to "pair up." For example, if those with

high college GPAs also tend to have high annual incomes (and those with low GPAs tend to have low annual incomes), the correlation coefficient would be *positive*. However, if people with high GPAs tend to have low annual incomes and those with low GPAs tend to have high annual incomes, the correlation coefficient would be *negative*. A coefficient's *strength* is revealed by its absolute value. For example, if a coefficient's absolute value is "close to" zero, the relationship between the two variables (i.e., their tendency to "pair up" positively or negatively) is weak and possibly due to sampling error. However, if a coefficient's absolute value is "far from" zero, the relationship between the two variables is strong and therefore unlikely to be due to sampling error. In this chapter, you will learn how to compute and then cautiously interpret correlation coefficients.

Reading Question

1. A correlation coefficient is used to
   a. compare the mean of one variable with the mean of another variable.
   b. determine if and how two variables might be related to each other.
   c. predict a person's score on one variable from his or her score on another variable.

Reading Question

2. A negative correlation coefficient means that
   a. if someone has a high score on one variable, he or she also tends to have a high score on the second variable.
   b. if someone has a low score on one variable, he or she also tends to have a low score on the second variable.
   c. if someone has a high score on one variable, he or she tends to have a low score on the second variable.

Reading Question

3. If two variables have a weak relationship, the absolute value of the correlation coefficient will be close to
   a. .05.
   b. 0.
   c. 1.

REVIEW OF *z* SCORES

Correlation coefficients are closely related to *z* scores; therefore, we will begin with a brief review of what you learned in Chapter 4. You learned that a *z* score enables you to locate individual scores relative to other scores in the same or different distributions. For example, if three people took a statistics exam, converting their raw scores to *z* scores would not only reveal who scored the highest but also how each person's score compared with the test's mean and if one or more of the scores were exceptionally high or exceptionally low. For instance, if Lola had a *z* score of +0.62, Bobby had a *z* score of −0.84, and Aaron had a *z* score of +1.8, you should know from this information that Lola scored a little better than average, Bobby scored a little worse than average, and Aaron scored exceptionally well.

Reading Question

4. Positive z scores represent values that are
   a. below average.
   b. average.
   c. above average.

Reading Question

5. The _____ of a z score indicates how much better or worse a given score is than the mean score.
   a. sign
   b. absolute value

THE LOGIC OF CORRELATION

The z score is called a univariate statistic because it uses data from a single variable, and conclusions drawn from the statistic must be limited to that variable. In the preceding example, the z scores allow us to comment on the relative performance of students on the statistics exam but nothing else. There are situations when knowing if scores on one variable are associated with scores on another variable is valuable. For example, if people who have higher depression scores also tend to have higher anxiety scores, this knowledge would help you understand both disorders more fully. Imagine that a group of people all completed a measure of depression, followed by a measure of anxiety. If you computed two z scores for each person, one for his or her depression score and one for his or her anxiety score, you could then compare individuals' two z scores and determine if people who had higher than average depression scores tended to have higher than average, average, or below-average anxiety scores. For example, Table 13.1 contains the depression and anxiety z scores for six people. The z scores are computed as $z = \frac{(X - M)}{SD}$. For Person A, the z score for depression is (44 – 60)/13.28 = −1.20, and the z score for anxiety is (59 – 79)/11.98 = −1.67.

An astute observer, with some knowledge of z scores, could probably look at the two columns of z scores above and conclude that, generally, those who have higher depression scores also tended to have higher anxiety scores, a positive relationship. While this general observation is true, there are at least two serious limitations with this "visual analysis." First, it would be much more difficult to discern this association if you had to examine pairs of z scores from 30, 60, or 100 people rather than just 6. Second, the general statement that if depression is high, then anxiety will be high is not very precise. It would be much more helpful if we (a) had a more reliable way of identifying if an association actually exists between depression scores and anxiety scores and (b) could quantify precisely how strong (or weak) the association is between depression and anxiety.

As you might have anticipated, correlation coefficients help us (a) determine if scores on two variables are associated with each other and (b) quantify the association's strength. The following correlation definitional formula reveals how z scores can quantify the association between two variables:

$$r = \frac{\sum Z_x Z_y}{N - 1}.$$

By looking at the above formula, you might be able to recognize that the correlation is the average product of each person's pair of z scores. In the depression and anxiety example, each person's

Table 13.1 Depression and Anxiety *z* Scores for 6 People

Person	*Depression Raw Score (X)*	*Depression z Score*	*Anxiety Raw Score (Y)*	*Anxiety z Score*
A	44	−1.20	59	−1.67
B	61	.08	85	.50
C	55	−.38	75	−.33
D	80	1.51	95	1.34
E	70	.75	82	.25
F	50	.75	78	−.08
	$M_x = 60$ $SD_x = 13.28$		$M_y = 79$ $SD_y = 11.98$	

depression and anxiety *z* scores are multiplied, and all the products are summed and then divided by the number of paired scores minus 1 ($N - 1$). In this case, the correlation would be computed as

$$r = \frac{\sum Z_x Z_y}{N-1}$$

$$r = \frac{(-.75)(-.08)+(.08)(.50)+(-.38)(-.33)+(1.51)(1.34)+(.75)(.25)+(-1.20)(-1.67)}{5} = .89.$$

As expected from your previous "visual analysis" of the *z* scores, the correlation is positive, indicating that depression and anxiety tend to coexist. In general, as depression levels increase, so do anxiety levels.

DIRECTION AND STRENGTH OF CORRELATION COEFFICIENTS

Correlation coefficients can vary between −1 and +1. As mentioned previously, the sign of the coefficient reveals the *direction* of the variables' relationship. A positive correlation reveals that the variables tend to have similar values (e.g., high depression scores pair with high anxiety scores, medium with medium, and low with low). A negative correlation reveals that the variables tend to have opposing values (e.g., high depression scores pair with low optimism scores and low depression scores pair with high optimism scores). The absolute value of the correlation reveals the *strength* of the relationship between the two variables. The more extreme the *r* value (i.e., the farther it is from 0), the stronger the relationship. A coefficient of 1 or −1 is the strongest association that is possible. Thus, the correlation of .89 indicates that the relationship between depression and anxiety is very strong.

Researchers frequently use scatterplot graphs to examine the relationship between variables. A trained eye can view a scatterplot and get a rough idea of both the direction and strength of the relationship between variables. For example, you can see that the relationship between depression and

anxiety is positive and strong by interpreting the scatterplot in Figure 13.1. Each point in a scatterplot represents one pairing of the variables being investigated. For example, the point at depression 44 and anxiety 59 indicates that one person in the data set had that combination of scores. When each participant's paired data point is included in the scatterplot, it reveals how the variables tend to "pair up." You can tell the relationship in Figure 13.1 is positive because the data points suggest a trend from the lower left (i.e., low depression scores and low anxiety scores) to the upper right (i.e., high depression scores and high anxiety scores). You can also tell that the relationship is quite strong because the data points form an "organized" line rather than a more haphazard cluster. The relationship in the first scatterplot in Figure 13.2 is the strongest possible positive correlation, +1. This relationship is a "perfect correlation" because the data points are perfectly organized; all of the data points fall "in line." Similarly, the second scatterplot in Figure 13.2 displays a perfect negative correlation because the data points are also perfectly organized into a straight line, but because the line trends from the upper left (i.e., low scores on the first variable and high scores on the other) to the lower right (i.e., high scores on the first variable and low scores on the other). In both scatterplots, the data points are perfectly organized into lines with no "scatter" from those lines, suggesting perfect relationships. As the data points in a scatterplot "scatter" from a perfectly organized line, the strength of the relationship between the two variables gets weaker.

For example, the scatterplots in Figure 13.3 display relationships that are more typical in research situations; variables rarely relate perfectly. You can see that while both graphs suggest trends, the individual data points do not form perfectly organized lines. The data points in both graphs are loosely organized into trends, but there is some "messiness," and therefore, these relationships are not perfect. The more the data points deviate from perfect organization, the weaker the relationship between the two variables. In the first scatterplot, in Figure 13.3, the data points suggest a moderately strong positive relationship because they form a moderately organized trend, from the lower left to the upper right. On the other hand, the data points in the second graph suggest a moderately strong negative relationship. If you compare the two graphs, you should be able to tell that the data points in the second graph are

Figure 13.1 Scatterplot of Depression and Anxiety Scores

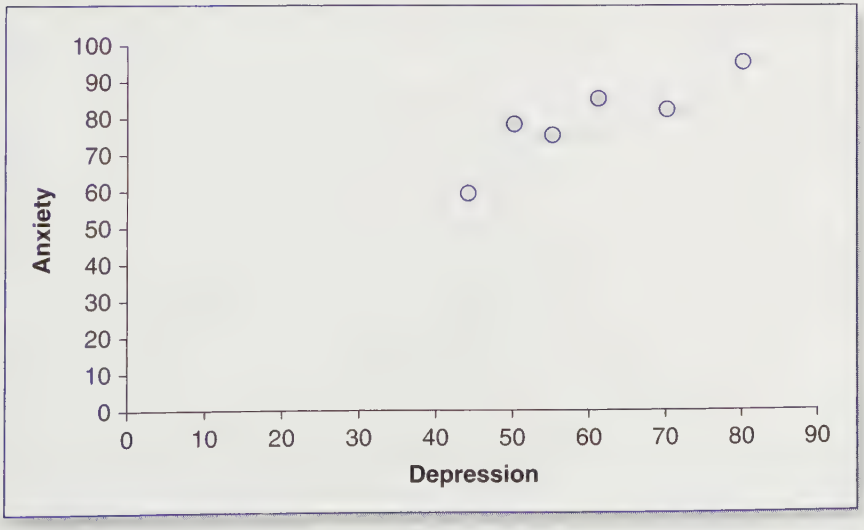


Figure 13.2 Perfect Positive and Negative Correlation Scatterplots

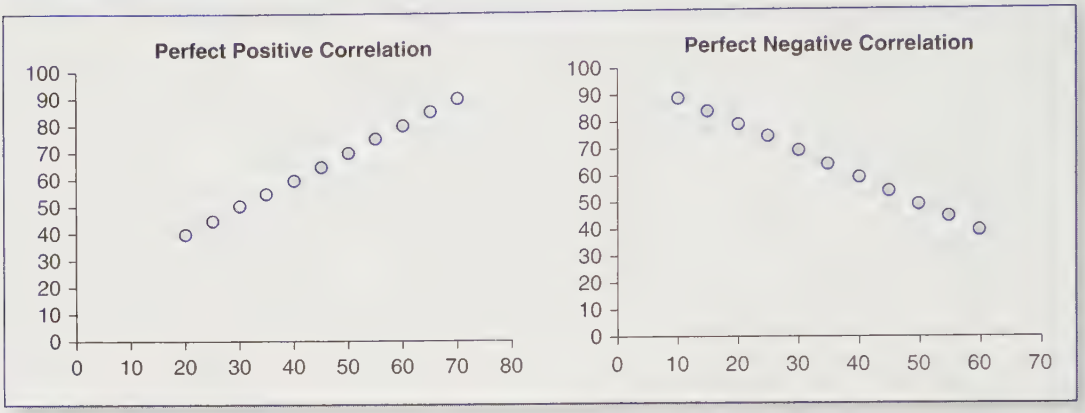
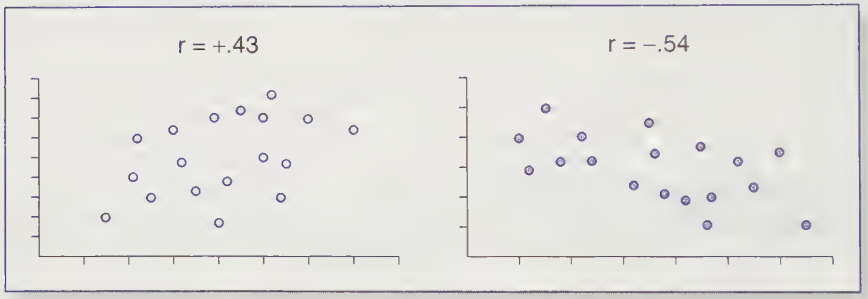


Figure 13.3 Scatterplots for Different Correlation Coefficients



"more organized" or "tighter" than the data in the first graph. This tighter organization represents a stronger relationship between the variables in the second graph than the first.

Reading Question

6. Which of the following are correlations designed to accomplish? Choose 2.
   a. Determine if two variables are associated
   b. Determine if a score is above or below the mean score
   c. Quantify the degree or strength of the association between two variables
   d. Quantify the distance that a score is above or below the mean score

Reading Question

7. When, on average, both paired z scores tend to be positive or both paired z scores tend to be negative, the resulting r value is
   a. negative.
   b. positive.

Reading Question

8. When, on average, paired z scores tend to have opposite signs, the resulting r value is
   a. negative.
   b. positive.

Reading Question

9. A negative correlation between two variables indicates that high scores on one variable are associated with _____ scores on the second variable.
   a. high
   b. low
   c. average

10. Which of the following indicates the strongest association between two variables?
    a. −.25
    b. .35
    c. −.67
    d. 1.1

COMPUTATIONAL FORMULAS

Definitional formulas help explain the logic of a statistic. The z score formula for correlation coefficients that we described at the beginning of the chapter is a definitional formula. It helps you understand how correlations work because it literally defines how the two types of scores are related. However, definitional formulas can be tedious to use. For example, if working with a large data set, it would be cumbersome to convert all raw scores to z scores and then compute an r value. **Computational formulas** are less intuitive, but they are easier to use. The following computational formulas allow you to compute an r value directly from raw scores without first converting everything to z scores; this saves a lot of time. While it is difficult to understand how computational formulas work by looking at them, they yield the exact same value as the definitional formula.

Reading Question

11. When computing statistics by hand, researchers most often use _____ formulas. When trying to understand how a statistic works, students will find it easier to work with _____ formulas. However, both formulas will always produce _____ value(s).
    a. computational; definitional; the same
    b. computational; definitional; different
    c. definitional; computational; the same
    d. definitional; computational; different

The computational approach requires you to use two formulas. You must first compute the shared variability of X and Y (i.e., SS_{xy}) and then use that value to compute r. Computing SS_{xy} requires $\sum XY$, $\sum X$, $\sum Y$, and N. $\sum XY$ is obtained by first multiplying the paired X and Y values and then summing the products. $\sum X$ and $\sum Y$ are the sum of the X and Y values, respectively. The sum-of-squared deviations (SS) for X and Y is then computed in the same way as in Chapter 3.

$$SS_{xy} = \Sigma XY - \frac{\Sigma X \Sigma Y}{N}.$$

$$r = \frac{SS_{XY}}{\sqrt{SS_X SS_Y}}.$$

Person	*Depression Raw Score (X)*	*Anxiety Raw Score (Y)*	XY
A	50	78	3,900
B	61	85	5,185
C	55	75	4,125
D	80	95	7,600
E	70	82	5,740
F	44	59	2,596
	$\Sigma X = 360$ $\Sigma X^2 = 22,482$	$\Sigma Y = 474$ $\Sigma Y^2 = 38,164$	$\Sigma XY = 29,146$

The numerator of the correlation (SS_{xy}) is computed as

$$SS_{xy} = \Sigma XY - \frac{\Sigma X \Sigma Y}{N} = 29,146 - \frac{(360)(474)}{6} = 706.$$

The denominator of the correlation is the square root of the product of SS_x and SS_y:

$$SS_x = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = 22,482 - \frac{360^2}{6} = 882.$$

$$SS_y = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N} = 38,164 - \frac{474^2}{6} = 718.$$

$$\sqrt{SS_x SS_y} = \sqrt{(882)(718)} = 795.79.$$

Finally, the r is

$$r = \frac{SS_{XY}}{\sqrt{SS_X SS_Y}} = \frac{706}{795.79} = .89.$$

This .89 correlation coefficient is identical to what you found when using the definitional formula at the beginning of the chapter.

SPEARMAN'S (r_s) CORRELATIONS

There are actually many different kinds of correlations. The one you just computed is called a Pearson's *r*. You should use Pearson's *r* when both variables are measured on interval or ratio scales. In the examples described earlier, you used Pearson's correlations because the variables of depression and anxiety were both measured on interval scales. If one or both of the variables are measured on an ordinal scale, you cannot use a Pearson's correlation. However, Spearman's correlation is very similar to Pearson's correlation, and it can analyze ordinal data. For example, you would use Spearman's *r* if you wanted to assess the relationship between the order of finish in a swimming race (first, second, third, etc.) and hours of swimming practice in the previous month. The computations for Spearman's correlation are identical to the computations for Pearson's correlation, with one important exception. The first step when computing Spearman's correlation is making both variables ordinal. For example, analyzing the relationship between finishing position in a race and hours of practice, you would have to convert the "hours of practice" variable into an ordinal variable. The person who practiced the most (e.g., 20 hours) would get a value of 1, the person with the next highest practice time (e.g., 17 hours) would get a 2, and so on. The "finishing position" variable is already an ordinal variable, so the next step is to do all the same computations you did for a Pearson's correlation but you analyze the two ordinal variables rather than the original pair of variables. Essentially, running a Spearman's correlation involves analyzing the rank orders of two variables rather than the variables themselves. If both variables are interval or ratio, you should use a Pearson's correlation by analyzing the raw data, but if either variable is ordinal, you need to use a Spearman's correlation by converting both variables to ranks and then analyzing those ranks.

Reading Question

12. If both variables being analyzed are measured on an interval or ratio scale, a _____ correlation should be used.
    a. Pearson's
    b. Spearman's

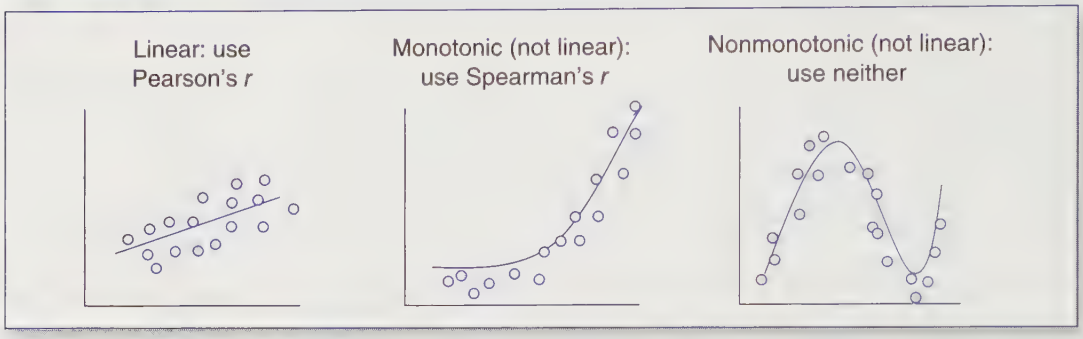
Reading Question

13. If one of the variables is measured on an ordinal scale and the other variable is measured on an ordinal, interval, or ratio scale, a _____ correlation should be used.
    a. Pearson's
    b. Spearman's

USING SCATTERPLOTS PRIOR TO CORRELATION COEFFICIENTS

Pearson's *r* is used when both variables are measured on interval or ratio scales. Before computing the correlation, you should first create a scatterplot to determine if the Pearson's correlation is the appropriate statistic to use. Pearson's correlation is only appropriate if there is a linear trend between the variables. A scatterplot of the two variables will reveal which correlation is appropriate. If, after creating a scatterplot, you discover that the data points do not follow a linear trend, you should not use Pearson's correlation. If the data are not linear but they are monotonic (see Figure 13.4), you can use the Spearman's correlation. *Monotonic* simply means that the data have a trend in only one direction

Figure 13.4 Graphs Representing Linear, Monotonic, and Nonmonotonic Relationships



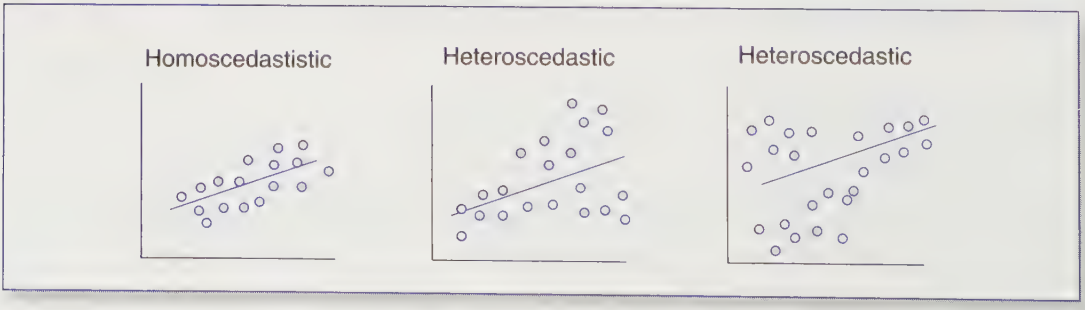
but not necessarily a linear trend. If the data trend upward and then downward, they are nonmonotonic (see Figure 13.4), and neither Pearson's nor Spearman's r can be used.

Reading Question

14. If the trend revealed by a scatterplot is not linear but it is monotonic, what correlation should be used?
    a. Pearson's
    b. Spearman's

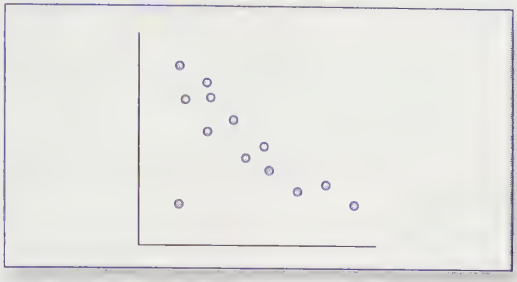
If the data form a linear trend, a Pearson's correlation might be appropriate, but you also need to look at the graph and check for homoscedasticity. Data are homoscedastic if the amount of variability in the data is similar across the entire scatterplot. For example, the first graph in Figure 13.5 shows a homoscedastic scatterplot. As the x variable (i.e., the one on the horizontal axis) increases, the variability around the "line of best fit" is about the same. However, the second and third graphs show data that are not homoscedastic but rather heteroscedastic. In these graphs, the data points become more variable as the x variable increases (second graph) or decreases (third graph). If the scatterplot suggests that the data are heteroscedastic, you should not use the Pearson's or Spearman's correlation.

Figure 13.5 Graphs Representing Heteroscedasticity and Homoscedasticity



When analyzing a scatterplot, you may find that a set of data are generally linear and homoscedastic, but they contain points that are outliers. For example, the graph in Figure 13.6 shows a strong, negative linear relationship, but there is one data point that deviates from the general trend. This one outlier can have a significant impact on the correlation. In these cases, researches usually look to determine if there is some sort of methodological or data entry error that led to this outlier. If you identify an error, fix it prior to analysis. If the outlying point is not an error, researchers often compute the correlation with and without the outlier and report both sets of analyses.

Figure 13.6 Graph of a Strong, Negative Linear Trend With an Outlier



ALTERNATIVE USE FOR CORRELATION

Most of the time when you use a correlation, you have scores from the same person for both variables. However, in some situations, you can assess the relationship between the scores from different individuals. For example, if you wanted to assess the relationship between spouses' marital satisfaction, you could measure the marital satisfaction of both people in a marriage to see if their respective satisfaction scores are associated. If one person in a marriage is satisfied, does the other person tend to be satisfied as well? And if so, how strong is the association between married couples' satisfaction scores? The key requirement is that the two variables being correlated must be "paired" in some way. The data must include paired scores that came from the same source (e.g., the same person, the same marriage).

Reading Question

15. A correlation can only be used if the scores on each variable
    a. come from the same people.
    b. are paired or linked to each other in some way.

CORRELATION AND CAUSATION

You have probably heard the common phrase, "correlation does not equal causation." Although the phrase is true, it is misleading. The popularity of this phrase, without similar cautions for other statistics (e.g., *t* tests, ANOVAs), incorrectly implies that these other statistics do allow you determine causal relationships. In fact, no statistic, by itself, allows researchers to infer causality. So, while it is true that correlation does not equal causation, it is also true that the "*t* test does not equal causation" and "ANOVA does not equal causation." To support a causal claim, there are two requirements. First, you must establish that the two variables are significantly related to each other. In other words, when one variable changes, the other variable changes as well. This first requirement is a statistical issue and can be addressed with *any* of the significance tests covered in this book. However, a statistically significant relationship is not sufficient evidence that one variable (IV) causes changes in the other (DV). The second requirement is that there are no confounds or alternative explanations for the

statistical association. As mentioned in other chapters of this text, eliminating confounds is a research methods issue. If you take a research methods course, you will learn how to control for confounds so you can eliminate alternative explanations. For now, you should understand that no statistic allows you to infer a causal relationship between an IV and a DV *unless* confounds are controlled. However, if confounds are controlled, *every* test statistic can support a causal conclusion, even correlation. So, the key point is "no statistical relationship equals causation." We know it doesn't rhyme like the more famous phrase, but it is more accurate.

Reading Question

16. Which of the following statements is true about statistics and causation?
    a. Correlations do not allow you to infer causality, but other statistics do.
    b. No statistics, on their own, are sufficient evidence for inferring causality.
    c. Correlations are not very useful to scientists because they do not enable researchers to make causal conclusions.

HYPOTHESIS TESTING WITH CORRELATION

As you know from your work with other statistics this semester, researchers often collect data from a representative sample and infer that the results they find approximate the results they would have found if they had studied the entire population (i.e., researchers often use inferential statistics). The same is done with both Pearson's and Spearman's correlations. In addition, as with the previous statistics you have worked with, researchers assume that a null hypothesis is true unless they find sufficient evidence to reject it. In the case of correlation, the null hypothesis is that the two variables being studied are not associated. If the null were true, the calculated r value would be close to 0. If the calculated r value is far from 0, the null is *not likely* to be true.

Reading Question

17. If the null hypothesis is true and two variables are not associated with each other, the r value should be close to
    a. −1.
    b. 1.
    c. 0.

As with other statistics, you use a critical value to define "far from 0." The critical value of r changes based on the sample size (i.e., $df = N - 2$). A table of critical r values is located in Appendix E. If the obtained r value is more extreme than the critical value (i.e., farther from 0), you should reject the null hypothesis. A complete example is provided below.

Reading Question

18. If an obtained r value is farther from 0 than the critical value, you should
    a. reject the null hypothesis.
    b. not reject the null hypothesis.

As with t tests, correlations can be one-tailed or two-tailed. If the research hypothesis predicts a positive association or a negative association, use a one-tailed test. If, however, the research hypothesis

does not predict a specific direction, only that the two variables will be related somehow, use a two-tailed test.

Reading Question

19. Use a two-tailed critical value when the research hypothesis
    a. predicts a positive correlation.
    b. does not predict a specific direction for the relationship between the two variables.
    c. predicts a negative correlation.

TWO-TAILED PEARSON'S CORRELATION EXAMPLE

You wonder if gratitude is associated with prosocial behavior. Specifically, you want to know if people who feel grateful about their own lives are more or less likely to want to help others. You do not understand these variables well enough to predict a positive or negative relationship; you just want to know if gratitude and prosocial beliefs are associated at all. To investigate this relationship, you have participants complete a questionnaire that measures gratitude and another that measures attitudes toward helping others. Both variables are measured on an interval/ratio scale and are computed such that higher scores indicate greater gratitude and stronger prosocial beliefs. The data from this study are given in Table 13.2. In this example, we will only use the computational formula for the r.

Table 13.2 Computation of a Correlation Coefficient Using the Definitional z Score Formulas

Participant	*Prosocial Attitudes (X)*	*Gratitude (Y)*	XY
A	1	3	3
B	3	4	12
C	3	4	12
D	3	6	18
E	1	4	4
F	3	5	15
G	4	6	24
H	2	2	4
I	3	4	12
J	3	6	18
	$\Sigma X = 26$	$\Sigma Y = 44$	$\Sigma XY = 122$

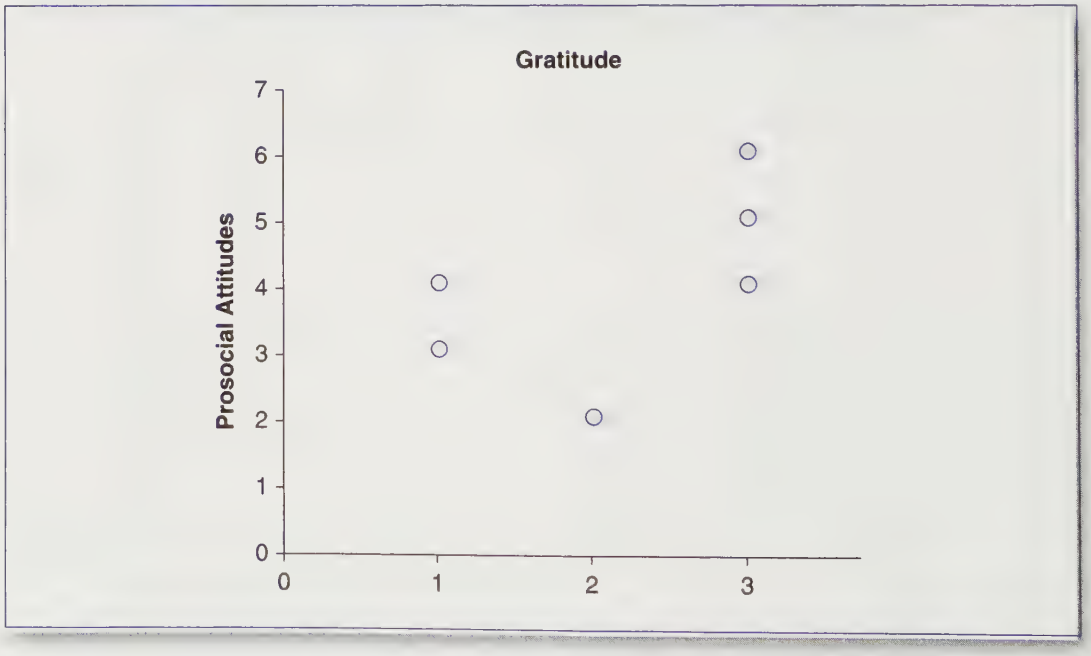
Step 1: Assess Statistical Assumptions

As with all statistical tests, the data collected from each participant must not be influenced by the responses of other participants. In this study, this *data independence assumption* is met. A Pearson's correlation requires that both variables be measured on an *interval/ratio scale of measurement*. In this study, both the measures of gratitude and prosocial attitudes met this assumption. If either variable were measured on an ordinal scale, you would need to run a Spearman's correlation. Pearson's correlation also requires that both variables have a normal shape in the population (*normality assumption*). Researchers typically assess this assumption by creating a histogram of each variable, and if the sample data have a normal shape, they assume that the population also has a normal shape.

The scatterplot in Figure 13.7 has the possible values for one variable on the x-axis (prosocial attitudes) and the possible values for the other variable on the y-axis (gratitude). It does not matter which variable goes on which axis. A dot at the paired coordinates represents each person's score on both variables. For example, Participant A's dot should be at the coordinates of prosocial attitude, 1, and gratitude, 3. Every person has a dot representing his or her combination of scores. If two or more people have the same scores for both variables, there will be two dots at the same coordinate location. The scores in Figure 13.7 seem to follow a straight line sloped upward, and so the trend represented in the figure is sufficiently linear to conduct a Pearson's correlation.

If the trend was not linear but still monotonic, you would use Spearman's correlation. If the trend was nonmonotonic, both the Pearson's and the Spearman's correlations would be inappropriate, and you should ask for help with your data.

Figure 13.7 Scatterplot of Gratitude and Prosocial Attitudes



Reading Question

20. In a scatterplot, each dot represents
    a. a set of paired X and Y scores.
    b. an X score.
    c. a Y score.

Step 2: State Null and Research Hypotheses Symbolically and Verbally

The research hypothesis is that the two variables have a linear relationship, whereas the null hypothesis is that they do not have a linear relationship. These hypotheses are presented verbally and symbolically in Table 13.3. You will notice that the Greek letter rho (ρ) represents the population parameter for correlations. In other words, the statistic r is an estimate of the population parameter (ρ).

Table 13.3 Symbolic and Verbal Representations of Two-Tailed Research and Null Hypotheses for a Pearson's Correlation

	Symbolic	*Verbal*	*Coefficient Created by*
Research hypothesis (H_1)	H_1: $\rho \neq 0$	Gratitude and prosocial attitudes are linearly related; the correlation is not zero.	Relationship between variables
Null hypothesis (H_0)	H_0: $\rho = 0$	Gratitude and prosocial attitudes are not linearly related; the correlation is zero.	Sampling error

Reading Question

21. The research hypothesis for a two-tailed Pearson's correlation predicts that the test statistic r will be far from
    a. 1.
    b. 0.
    c. −1.

Step 3: Define the Critical Region

The degrees of freedom formula for correlations is $df = N - 2$, where N is the number of paired scores, not the number of scores. Therefore, in this case,

$$df = N - 2 = 10 - 2 = 8.$$

The two-tailed critical value associated with $df = 8$ can be found in the table of Pearson's r critical values in Appendix E. The two critical regions for this example are +.632 and −.632. If the obtained r value is more extreme than these values, the null hypothesis should be rejected. If the obtained r is in the negative tail of the r distribution, the two variables have a significant negative association, and if it is in the positive tail, they have a positive association.

Reading Question

22. A two-tailed correlation has _____ critical region(s).
    a. no
    b. one
    c. two
    d. three

Step 4: Compute the Test Statistic (Pearson's *r*)

Some of the computational work was completed in the above data table.

The numerator of the correlation (SS_{xy}) is computed as

$$SS_{xy} = \sum XY - \frac{\sum X \sum Y}{N} = 122 - \frac{(26)(44)}{10} = 7.6$$

The denominator of the correlation is the square root of the product of SS_x and SS_y:

$$SS_x = \sum X^2 - \frac{(\sum X)^2}{N} = 76 - \frac{26^2}{10} = 8.4.$$

$$SS_y = \sum Y^2 - \frac{(\sum Y)^2}{N} = 210 - \frac{44^2}{10} = 16.4.$$

$$\sqrt{SS_x SS_y} = \sqrt{(8.4)(16.4)} = 11.74.$$

Finally, the *r* is

$$r = \frac{SS_{XY}}{\sqrt{SS_X SS_Y}} = \frac{7.3}{11.74} = .65.$$

The obtained *r* value of .65 was in the positive critical region (i.e., it was farther from 0 than the critical value of +.632), which means that gratitude and prosocial attitudes have a significant positive association. That is, those who are more grateful also seem to have more positive attitudes toward helping others.

Step 5: Compute the Effect Size (r^2) and Describe It

The size of a correlation is described by r^2, which is also called the "coefficient of determination."

$$r^2 = (.65)^2 = .42.$$

The coefficient of determination is interpreted as a percentage. Specifically, $r^2 = .42$ indicates that 42% of the variability in prosocial attitudes is predicted by the variability in gratitude scores. If the r^2 between prosocial attitudes and gratitude were 1, you could perfectly predict someone's prosocial attitudes from his or her gratitude. The general guidelines for interpreting r^2 are presented in Table 13.4 and are the same as those for η_p^2.

Reading Question

23. If $r^2 = .36$, it means that 36% of the variability in one variable is
    a. unpredictable.
    b. predicted by the variability in the other variable.

Table 13.4 General Guidelines for Interpreting r^2

r^2 Value	Effect Size Label
Close to .01	Small
Close to .09	Medium
Close to .25	Large

Step 6: Summarize the Results

This correlation analysis can be summarized by the following sentences.

There is a positive association between gratitude and prosocial attitudes, $r(8) = .65$, $p < .05$. That is, those with higher gratitude scores also tended to have more positive attitudes toward helping others.

The APA reporting format for correlations is similar to that used for other statistics; however, with other statistics, an effect size estimate is included in the summary statement. This is typically not done with correlations because the effect size estimate is so easily computed from the r value, which is included in the APA reporting format (i.e., effect size = r^2).

Reading Question

24. If $r(8) = .65$, $p < .05$, the number of paired scores in the study was _____ and r^2 = _____.
    a. 10; .65
    b. 8; .4225
    c. 8; .65
    d. 10; .4225

ONE-TAILED PEARSON'S CORRELATION EXAMPLE

Now, you are interested in replicating your previous work on gratitude and prosocial attitudes at your college. In your previous study, you found a strong positive association, and you expect to find a similar relationship at your college. Therefore, your research hypothesis is that there will be a positive linear correlation between these variables. You collect data from 10 students attending your college. The data are presented in Table 13.5.

Step 1: Assess Statistical Assumptions

As was the case in your previous study, all of the statistical assumptions are met. You then create a scatterplot to determine whether a Pearson's or Spearman's correlation is appropriate. The data for this example are plotted in Figure 13.8. The trend represented in the figure is sufficiently linear to conduct a Pearson's correlation.

Step 2: State the Null and Research Hypotheses Symbolically and Verbally

Given the results of your previous study, you are predicting a positive correlation between gratitude and prosocial attitudes, so you correctly choose to do a one-tailed significance test. The null and research hypotheses are presented in Table 13.6.

Table 13.5 Computation of a Correlation Coefficient Using the Computational Formulas

Participant	*Prosocial Attitudes (X)*	*Gratitude (Y)*	XY
A	1	2	2
B	2	2	4
C	2	4	8
D	3	4	12
E	1	3	3
F	3	3	9
G	2	4	8
H	2	3	6
I	3	3	9
J	3	4	12
	$\sum X = 22$ $SS_x = 5.6$	$\sum Y = 32$ $SS_y = 5.6$	$\sum XY = 73$

Figure 13.8 Scatterplot of Scores for Prosocial Attitudes and Gratitude

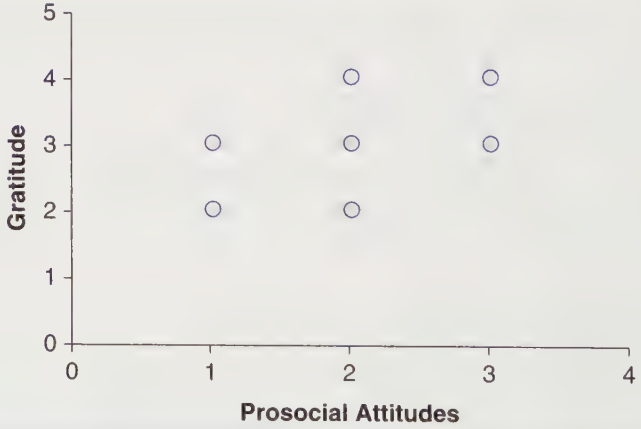


Table 13.6 Symbolic and Verbal Representations of the One-Tailed Research and Null Hypotheses for a Pearson's Correlation

	Symbolic	*Verbal*	*Coefficient Created by*
Research hypothesis (H_1)	H_1: $\rho > 0$	Gratitude and prosocial attitudes are positively related.	Relationship between variables
Null hypothesis (H_0)	H_0: $\rho \leq 0$	Gratitude and prosocial attitudes are not positively related.	Sampling error

Step 3: Define the Critical Region

In this case, $df = N - 2 = 10 - 2 = 8$.

The one-tailed critical value associated with a $df = 8$ is .549. You predicted a positive correlation, so the critical region for this example is +.549. If the obtained r value is larger than this critical r value, you should reject the null hypothesis.

Step 4: Compute the Test Statistic (Pearson's r)

You use the computational formula to first compute SS_{xy} and then r.

$$SS_{xy} = \sum XY - \frac{(\sum X)(\sum Y)}{N} = 73 - \frac{(22)(32)}{10} = 2.6.$$

$$r = \frac{SS_{XY}}{\sqrt{(SS_X)(SS_Y)}} = \frac{2.6}{\sqrt{(5.6)(5.6)}} = .46.$$

The obtained r value of .46 was not in the critical region, which means that gratitude and prosocial attitudes were not significantly positively correlated in your sample.

Step 5: Compute the Effect Size (r^2) and Describe It

In this case, r^2, or the coefficient of determination, was $.46^2$, or .21. This indicates a medium to large, positive association between gratitude and prosocial attitudes. Although the obtained r value was not sufficient to reject the null hypothesis, the medium to large effect size suggests that the two variables may be associated but the sample size was too small to detect the relationship between the variables.

Step 6: Summarize the Results

The following short paragraph summarizes your results.

The students' gratitude scores and prosocial attitudes were not significantly correlated, $r(8) = .46$, $p > .05$. However, the medium to large association between the variables suggests that the null hypothesis may not have been rejected because the sample size ($N = 10$) was too small. A larger sample size is needed to study the relationship between these variables.

Reading Question

25. In an APA-style reporting statement, the number in parentheses after r is
    a. the Pearson's correlation.
    b. the degrees of freedom.

WHAT IF YOU NEED TO DO A SPEARMAN'S CORRELATION?

You learned that if a scatterplot is monotonic rather than linear, you must do a Spearman's correlation rather than a Pearson's correlation. The interesting fact about Spearman's correlation is that it uses the exact same formulas as Pearson's correlation. The only difference between the two correlations is in the type of data that are used. While Pearson's correlation uses the raw scores, *Spearman's correlation analyzes the ranks of the scores* rather than the scores themselves. Therefore, if you need to conduct a Spearman's correlation by hand, you must first convert all of the scores for Variable 1 into ranks and all of the scores for Variable 2 into ranks.

Table 13.7 displays how the raw scores for each variable would be converted into ranks for a small data set of four scores for each variable. Higher scores represent better scores for both Variable 1 and Variable 2.

After converting the raw scores for Variables 1 and 2 into ranks, the remaining steps for computing Spearman's correlation are the same as those for Pearson's correlation. If you are using SPSS to conduct a Spearman's correlation, the program will convert the raw scores to ranked scores for you. The only thing you will have to do is indicate that you want a Spearman's rather than a Pearson's correlation.

Table 13.7 Converting Raw Scores Into Rank Scores for Spearman's Correlation

Variable 1 Raw Scores	*Variable 1 Rank Scores*	*Variable 2 Raw Scores*	*Variable 2 Rank Scores*
23	3	10	1
15	4	5	3
30	1	9	2
27	2	2	4

Reading Question

26. If you need to do a Spearman's correlation because the data are monotonic but not linear, you need to convert the raw scores to _____ and then use the same steps/formulas that you would use to perform a Pearson's correlation.
    a. z scores
    b. ranked scores

CONFIDENCE INTERVALS

In previous chapters, we computed confidence intervals around means and mean differences. You can also compute confidence intervals around correlations to determine a range of plausible values for the value of the correlation in the population. The confidence interval formula for correlations has the

same general format as for all other confidence intervals. You begin with a point estimate, which is the r you calculate, and then you subtract the margin of error to obtain the lower bound of the confidence interval and add the margin of error to obtain the upper bound.

$$\text{Lower Bound} = \text{Point estimate of } r - \text{Margin of Error}$$

$$\text{Upper Bound} = \text{Point estimate of } r + \text{Margin of Error}$$

Although the format of the confidence interval is identical to the confidence intervals you computed around means and mean differences, there are some differences in the computations. These differences are discussed in Activity 13.2.

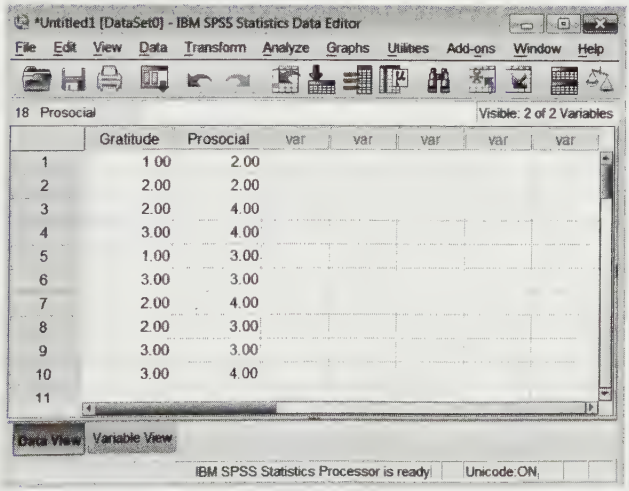
SPSS

In this section, we reanalyze the last example in this chapter to illustrate how to use SPSS to perform a correlation.

Data File

After you enter your data, it should look like the screenshot in Figure 13.9.

Figure 13.9 SPSS Screenshot of Data Entry Screen



Reading Question

27. When entering data for a correlation analysis in SPSS,
    a. each variable is in its own column.
    b. each set of paired scores is in its own column.

Creating a Scatterplot

There are two ways to create scatterplots in SPSS:

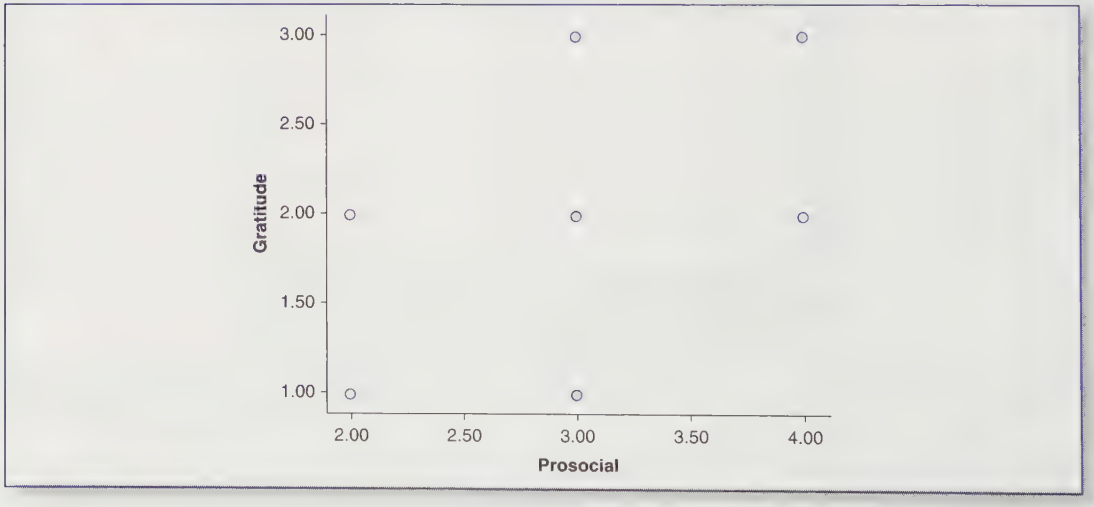
1. Option 1(Legacy Dialogs)
   - Go to the Graphs menu. Choose Legacy Dialogs, and then select Scatter/Dot.
   - Choose Simple Scatter, and click Define.
   - Put one variable on the x-axis and the other variable on the y-axis.

2. Option 2 (Chart Builder)
   - Go to the Graphs menu, and select Chart Builder.
   - A box may pop up that warns you to make sure that the appropriate level of measurement is specified in the data file. Close this box by clicking on OK.
   - You will see two windows. One is labeled "Element Properties," and the other is labeled "Chart Builder." Close the Element Properties window.
   - To create the scatterplot, click on the Gallery tab on the lower half of the screen, and then click on "Scatter/Dot."
   - Once you have done that, there will be several pictures of different types of scatterplots. Double-click on the first one. If you hold your cursor over it, it will say "Simple Scatter."
   - Next, you need to indicate which variable you want on the x-axis and which you want on the y-axis. Just choose one variable from the Variable box and drag it onto the x-axis, and then drag the other variable onto the y-axis. Click OK to create the scatterplot.

Output

Your output file should look similar to the one in Figure 13.10. (Note: If you put gratitude on the horizontal axis, your dots will be in different locations.) The plot in Figure 13.10 is sufficiently linear to conduct a Pearson's correlation.

Figure 13.10 SPSS Screenshot of Scatterplot



Reading Question

28. Which of the following sets of paired scores is not represented in the above scatterplot? The prosocial score is listed first, followed by the gratitude score.
    a. 2, 2
    b. 3, 2
    c. 4, 6
    d. 3, 3

Computing a Correlation

Go to the Analyze menu. Choose Correlate, and then select Bivariate.

Move the variables you want a correlation for into the Variables box.

Check Pearson's or Spearman's correlation.

The output you see in Figure 13.11 is called a correlation matrix. Note that the correlation, or r value, of .648 is the same as the r value that was computed by hand at the beginning of the chapter. Also, note that SPSS always presents the correlation twice. The Sig. (2-tailed) value, or p value, for this correlation is .043. This means that when the sample size is $N = 10$, an r value of .648 would be expected to occur about four times out of 100 due to sampling error. When interpreting the results from the computer output, you compare the p value (i.e., .043) with the alpha value that was set at the beginning of the study; in this case, $\alpha = .05$. Because the p value is less than the alpha value, that is, p (.043) < α (.05), you should reject the null hypothesis. You should recognize that this is the same conclusion that you reached when you analyzed these data by hand.

Figure 13.11 SPSS Screenshot of Correlation Output

Correlations

		Gratitude	Prosocial
Gratitude	Pearson Correlation	1	.648*
	Sig. (2-tailed)		.043
	N	10	10
Prosocial	Pearson Correlation	.648*	1
	Sig. (2-tailed)	.043	
	N	10	10

*. Correlation is significant at the 0.05 level (2-tailed).

Reading Question

29. What is the exact p value produced by the SPSS output on page 535?
    a. 1
    b. 10
    c. .648
    d. .043

OVERVIEW OF THE ACTIVITIES

In Activity 13.1, you will learn to estimate the strength and direction of correlations from scatterplots, how to recognize linear versus nonlinear trends in scatterplots, and work through the six steps of hypothesis testing for a correlation with "hand" calculations and SPSS. Activity 13.2 has you compute confidence intervals for correlations. In Activity 13.3, you will compute a Spearman's correlation. Activity 13.4 introduces the statistical procedure of linear regression. Finally, Activity 13.5 has you choose which statistical procedures are appropriate for various research situations.

Activity 13.1: Correlations

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Determine if it is appropriate to use Pearson's or Spearman's correlation by viewing a scatterplot
- Estimate the correlation coefficient by inspecting a scatterplot
- Conduct a one-tailed and a two-tailed hypothesis test using a Pearson's correlation
- Compute a Pearson's correlation by hand and using SPSS
- Summarize the results of a correlation in APA format

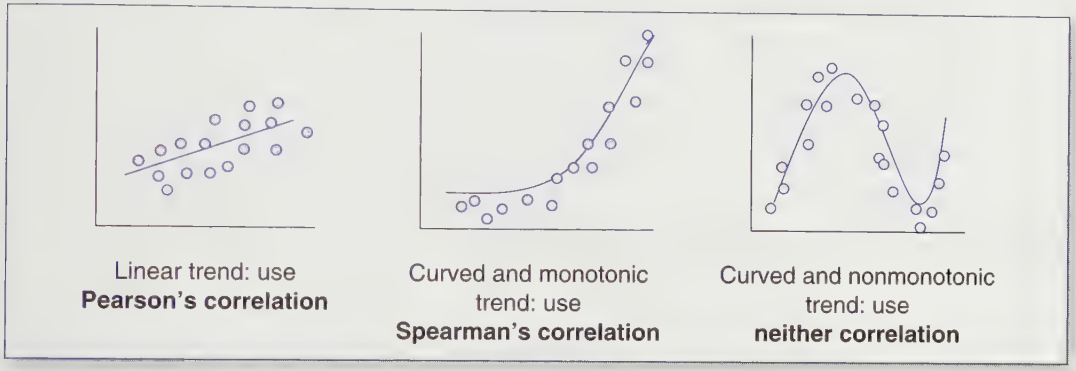
PART I: JUDGING CORRELATION STRENGTH FROM SCATTERPLOTS

Before computing correlations, you should get some practice looking at scatterplots and determining the strength of the relationship between two variables just based on the graph. In general, the closer the data points are to a straight line, the stronger the relationship between the variables. Go to the following website, and answer the questions until you can accurately associate a graph with the appropriate correlation coefficient: http://istics.net/stat/Correlations/.

1. My longest correct answer streak was _____.
2. What characteristic of a scatterplot determines the strength of the relationship between the two variables? Describe this characteristic to someone who has not taken a statistics class.
3. What characteristic of a scatterplot determines the direction of the relationship between the two variables? Describe this characteristic to someone who has not taken a statistic class.

PART II: IDENTIFYING APPROPRIATE CORRELATION BY VIEWING SCATTERPLOTS

Before computing the correlation between two quantitative variables, you should generate a scatterplot of the data to determine if Pearson's or Spearman's correlation, or neither of them, should be used. As discussed in the reading, if the scatterplot is *linear,* use a Pearson's correlation. If the scatterplot is *monotonic* or "slightly curvy but not dramatically changing its overall trend," use a Spearman's correlation. And if the scatterplot is *nonmonotonic* or "dramatically changing directions," use neither Pearson's nor Spearman's correlations. The figures exemplify possible trends and the type of correlation that should be used to analyze them.



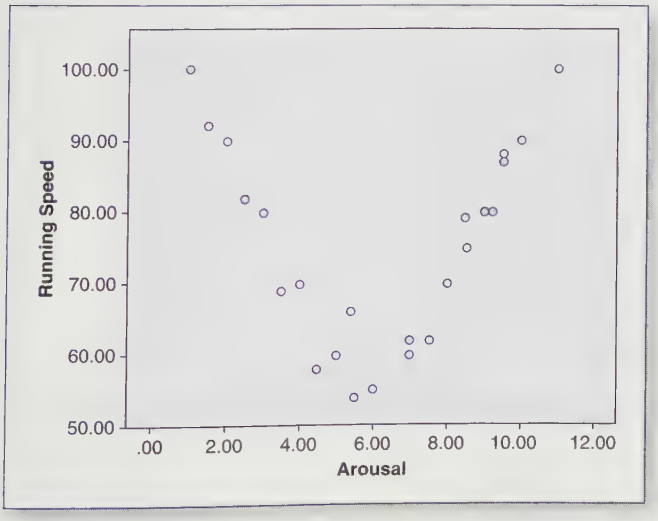
Linear trend: use **Pearson's correlation**

Curved and monotonic trend: use **Spearman's correlation**

Curved and nonmonotonic trend: use **neither correlation**

For the next four questions, you should look at the scatterplot and determine if a Pearson's or Spearman's correlation should be performed or neither of the two.

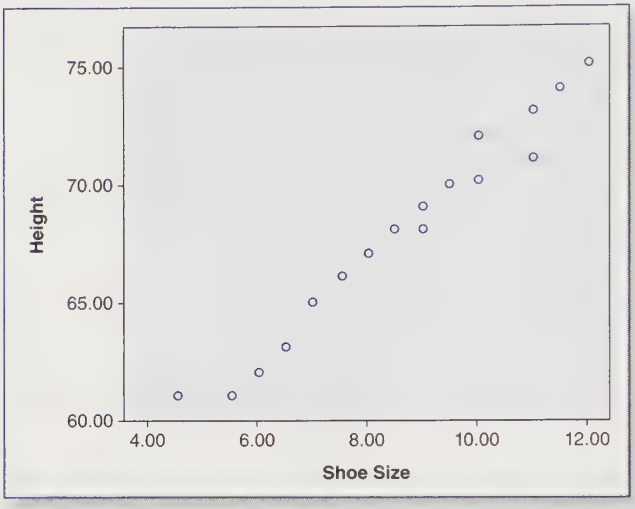
4. To determine if arousal is associated with running speed, a sports psychologist obtains a sample of runners and measures their reported level of arousal on a 12-point scale, with higher numbers indicating higher levels of arousal. After measuring their arousal, each person is asked to run 400 m, and the times are recorded.



Based on the scatterplot, which correlation is appropriate for these data?

a. Pearson's
b. Spearman's
c. Neither

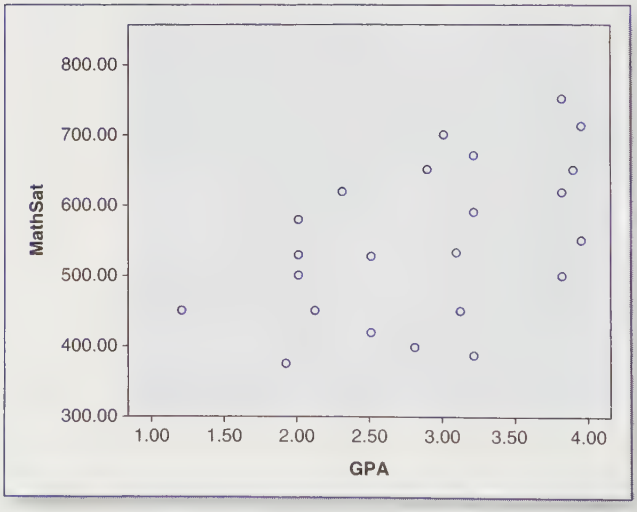
5. The following height and shoe size data came from a sample of 17 college students:



Based on the scatterplot, which correlation would be appropriate for these data?

a. Pearson's
b. Spearman's
c. Neither

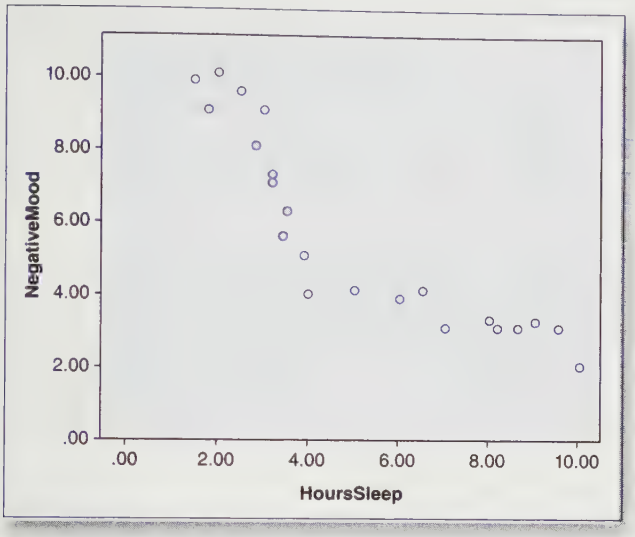
6. The members of a college admissions committee want to know if the Math portion of the SAT is a good predictor of college performance. They obtain Math SAT scores as well as first-year college GPAs from a sample of 23 students.



Based on the scatterplot, which correlation is appropriate for these data?

a. Pearson's
b. Spearman's
c. Neither

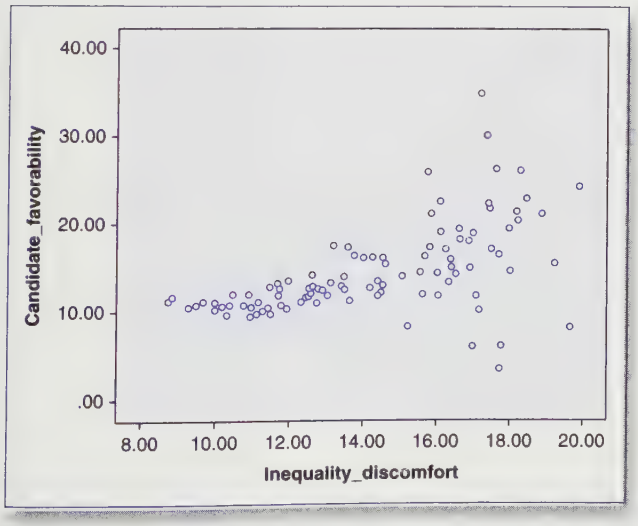
7. A student notices that she is usually in a bad mood when she is sleep deprived and wonders if this is true for other people. To investigate this, she asks students to report the number of hours of sleep they received the previous night and conducts a survey assessing their current mood (higher numbers indicate a more negative mood).



Based on the scatterplot, which correlation is appropriate for these data?

a. Pearson's
b. Spearman's
c. Neither

8. A researcher wants to test the relationship between a scale that measures discomfort with income inequality and a scale that measures how much they like a specific candidate for president.



Based on the scatterplot, which correlation is appropriate for these data?

a. Pearson's
b. Spearman's
c. Neither

PART III: A RESEARCH EXAMPLE

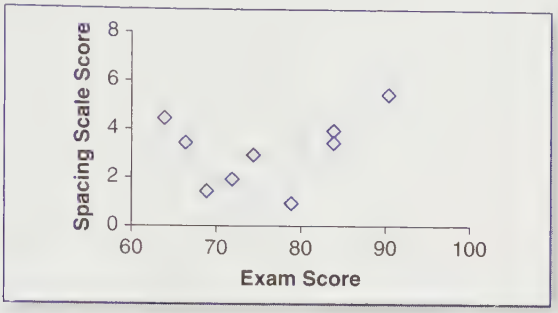
Dunlosky, Rawson, Marsh, Nathan, and Willingham (2013) reviewed over 700 scientific articles on 10 learning strategies commonly used by college students. Based on their review, they highly recommended two studying strategies because the research suggested that they were easy to use and that they were highly effective in a wide range of classes. Specifically, they recommended that students study by using **practice retrieval** (i.e., self-testing) and spreading studying across days rather than cramming (i.e., spacing effect or **distributed practice**). They also found that the studying strategy that was most commonly used by college students, **rereading chapters**, was *not* a consistently effective studying strategy. Your research team wants to determine if these recommendations are supported by data collected at your university. Specifically, your team wants to know if using practice retrieval and distributed practice is associated with higher exam scores and if rereading chapters is actually an *in*effective strategy as Dunlosky et al. (2013) suggest.

The first step in your research is to run a pilot study. Your team collects a small sample of 10 students currently taking an introductory psychology course. You ask them to complete a questionnaire about their studying habits. Some of your questions assessed the degree to which the students used practice retrieval when studying (e.g., "I tried to recall key terms or concepts from memory while studying"). Other questions assessed the degree to which the students used distributed practice or spacing (e.g., "I spread out my studying for the exam across multiple days"). Still other questions assessed the degree to which the students used rereading as a studying strategy, (e.g., "To study for the exam, I reread all of the chapters"). Your sample of students responded to all of these questions by using a 7-point scale with 1 = *never* and 7 = *every time*. Your team combined certain questions to create a Practice Retrieval Scale, a Spacing Scale, and a Rereading Scale. Higher scores on any of these scales reflect more use of that studying strategy. Finally, you also collected students' exam scores.

Your team divided up the work. You are responsible for analyzing the correlation between Spacing Scale scores and exam scores. The data are presented below. These data do meet all of the necessary statistical assumptions.

Participant	*Exam Score (X)*	*Spacing Scale Score (Y)*
A	69	1.5
B	84	4
C	90.5	5.5
D	72	2
E	84	3.5
F	79	1
G	74.5	3
H	66.5	3.5
I	64	4.5
J	77	2.5

9. Your first step is inspecting the scatterplot of the exam score and Spacing Scale score. Most of the scatterplot is created for you below, but the last data point from Participant J is missing. Place that data point into the scatterplot.



10. Do the sample data look like they are linear enough to perform a Pearson's correlation? (Generally, a scatterplot with only 10 data points doesn't really give you enough information to be confident, so unless it is obviously nonlinear, you should proceed with doing a Pearson's correlation.)

    a. Yes

    b. No

11. Based on the predictions derived from the Dunlosky et al. (2013) study, place an H_0 next to the one-tailed null hypothesis and an H_1 next to the one-tailed research hypothesis.

    _____ Higher Spacing Scale scores will be associated with higher exam scores.
    _____ Lower Spacing Scale scores will be associated with higher exam scores.
    _____ Higher Spacing Scale scores will be associated with lower exam scores.
    _____ Higher Spacing Scale scores will not be associated with higher exam scores.
    _____ Lower Spacing Scale scores will not be associated with higher exam scores.
    _____ Higher Spacing Scale scores will not be associated with lower exam scores.

12. Compute the *df* and define the critical region for *r;* use $\alpha = .05$.

13. Use the computational formula to compute the correlation coefficient by hand. All the preliminary computations are done for you below. Take a few moments to make sure you understand how each value was computed.

Participant	*Exam Score (X)*	X^2	*Spacing Scale (Y)*	Y^2	X * Y
A	69	4,761	1.5	2.25	103.5
B	84	7,056	4	16	336
C	90.5	8,190.25	5.5	30.25	497.75
D	72	5,184	2	4	144

(Continued)

(Continued)

Participant	*Exam Score (X)*	X^2	*Spacing Scale (Y)*	Y^2	X * Y
E	84	7,056	3.5	12.25	294
F	79	6,241	1	1	79
G	74.5	5,550.25	3	9	223.5
H	66.5	4,422.25	3.5	12.25	232.75
I	64	4,096	4.5	20.25	288
J	77	5,929	2.5	6.25	192.5
	$\Sigma X = 760.5$	$\Sigma X^2 = 58{,}485.75$	$\Sigma Y = 31$	$\Sigma Y^2 = 113.5$	$\Sigma XY = 2{,}391$

$$SS_{xy} = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N} =$$

$$SS_X = \Sigma X^2 - \frac{(\Sigma X)^2}{N} =$$

$$SS_Y = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N} =$$

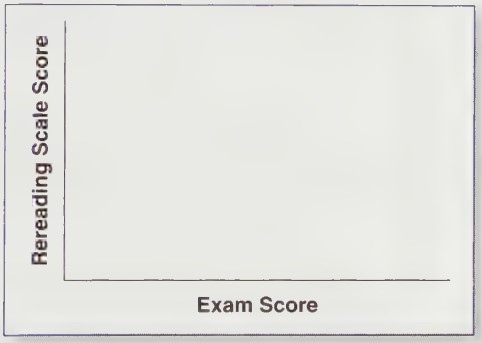
$$r = \frac{SS_{XY}}{\sqrt{(SS_X)(SS_Y)}} =$$

14. Compute the effect size (r^2) for this study.
15. How large in the effect? (small, small to medium, etc.)
16. Choose which of the following paragraphs best summarizes the correlation results. Then fill in the blanks with the appropriate statistical information.
    a. There was a significant correlation between the Spacing Scale scores and exam scores of the students, r(____) = _____, $p < .05$. Those with high Spacing Scale scores tended to have higher exam scores.
    b. There was no significant correlation between the Spacing Scale scores and exam scores of the students, r(____) = _____, $p > .05$. However, the effect size of the correlation was medium, so the study should be redone with a larger sample size.
    c. There was a significant correlation between the Spacing Scale scores and exam scores of the students, r (____) = _____, $p < .05$.

Now you have to compute the correlation between the students' exam scores (X) and Rereading Scale scores (Y). The data are below. Some of the preliminary computations are also provided.

Participant	*Exam Score (X)*	X^2	*Rereading (Y)*	Y^2	X * Y
A	69	4,761	5		
B	84	7,056	7		
C	90.5	8,190.25	4		
D	72	5,184	5.5		
E	84	7,056	4.5		
F	79	6,241	4		
G	74.5	5,550.25	4.5		
H	66.5	4,422.25	5		
I	64	4,096	4		
J	77	5,929	6		
	$\Sigma X = 760.5$	$\Sigma X^2 = 58{,}485.75$	$\Sigma Y =$	$\Sigma Y^2 =$	$\Sigma XY =$

17. Create a scatterplot of the relationship between exam score and Rereading Scale score.



18. Choose the two-tailed null hypothesis for this study.

    a. $\rho = 0$

    b. $\rho \neq 0$

19. Compute the *df* and define the critical region for *r;* use $\alpha = .05$.

20. Use the computational formula to compute the correlation coefficient. Some of the preliminary computations are done for you in the above table.

$$SS_{xy} = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N} =$$

$$SS_X = \Sigma X^2 - \frac{(\Sigma X)^2}{N} =$$

$$SS_Y = \sum Y^2 - \frac{(\sum Y)^2}{N} =$$

$$r = \frac{SS_{XY}}{\sqrt{(SS_X)(SS_Y)}} =$$

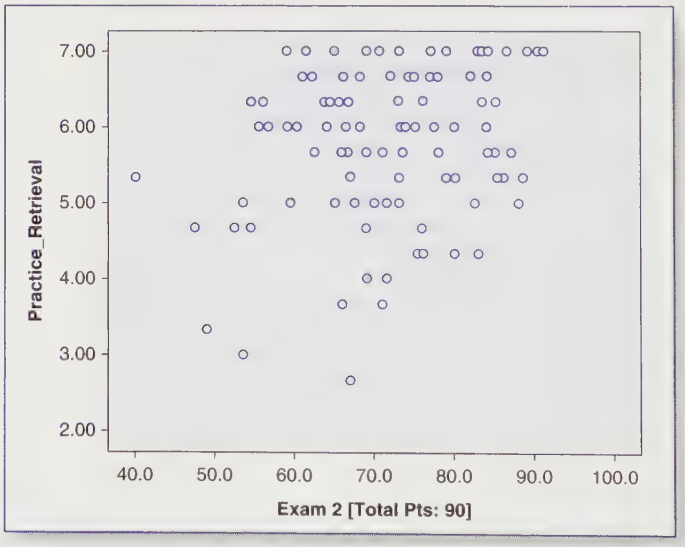
21. Compute the effect size (r^2) for this study, and interpret it as small, small to medium, and so on.

22. Write an APA-style summary of the results. Use the examples from Question 16 as a guide.

Your team used the pilot study to fine-tune the data collection procedure. Now you are ready to conduct the larger study. Again, your research goal is to determine if practice retrieval and spacing studying strategies are each positively correlated with exam scores. You also want to know if rereading is an ineffective studying strategy. Your team had all students in an introductory psychology course complete your questionnaire on studying habits so you now have practice retrieval scores, spacing scores, and rereading scores for every student. You also have each student's exam score. Someone on your research team entered all of this data into the file, "Correlation Studying Strategies.sav." Use SPSS to analyze these data.

As your first step in the analysis, you should examine the scatterplots for each of the three correlations you are interested in computing—namely, the scatterplot for (1) exam score and practice retrieval, (2) exam score and spacing, and (3) exam score and rereading. These scatterplots follow, but you should use SPSS to generate them for yourself.

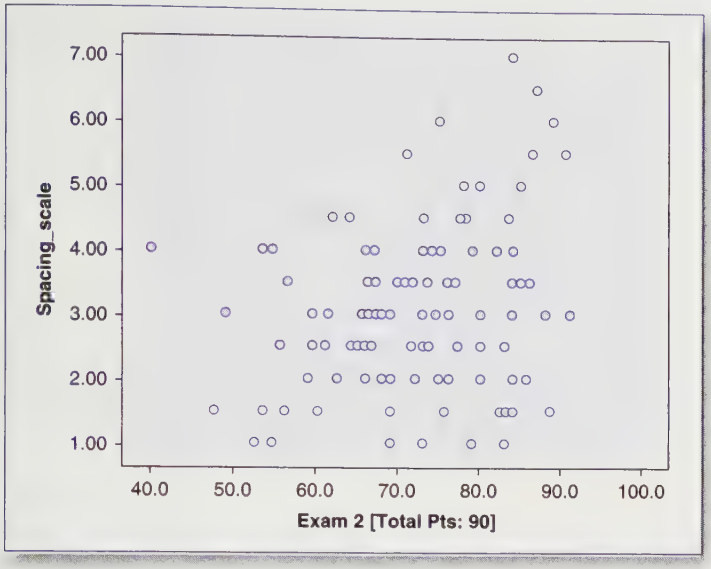
- Clicking on Graphs → Legacy Dialogues → Scatter/Dot . . . → Simple Scatter → Define.
- Move Exam_Score onto the *x*-axis and Practice_Retrieval onto the *y*-axis. Click OK.
- Repeat the same steps for the other two scatterplots. Make sure the graphs you get look like those provided below.



23. Does this relationship look linear? In other words, can you run a Pearson's correlation on these data?

    a. Yes

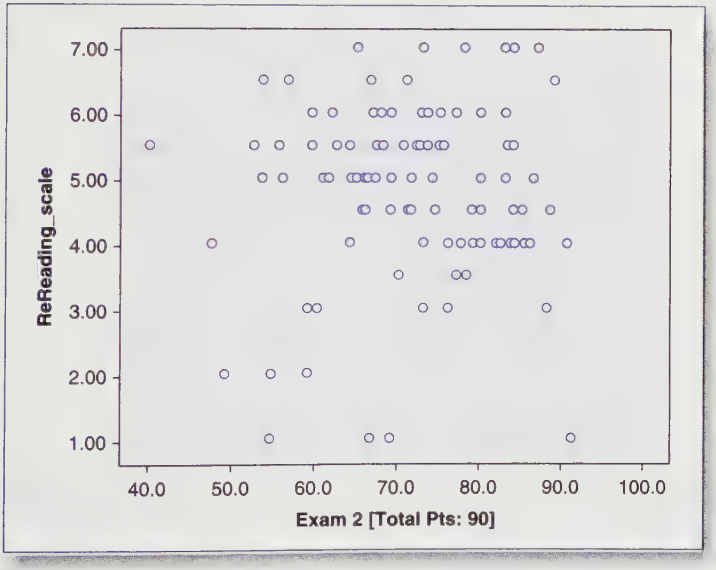
    b. No



24. Does this relationship look linear? In other words, can you run a Pearson's correlation on these data?

    a. Yes

    b. No



25. Does this relationship look linear? In other words, can you run a Pearson's correlation on these data?

    a. Yes

    b. No

26. Now you need to use SPSS to run these three correlations. SPSS can generate all three correlations at the same time and provide you with a "correlation matrix." A correlation matrix is a table that displays how each variable is correlated with each of the other variables.

    - Click on Analyze → Correlate → Bivariate.
    - Move all of the variables into the Variables box.
    - Make sure the Pearson correlation box is checked.

- You can also tell SPSS to provide one- or two-tailed p values by choosing the one you want in the "Test of Significance" box. Obtain the correlation matrix from SPSS.

27. Inspect the correlation matrix you obtained. It provides more than the three correlations you need to test the three predictions of Dunlosky et al. (2013). How many *unique* correlations are in the correlation matrix you obtained? _____

28. Find the three correlations you need to test Dunlosky et al.'s predictions; write down the *df*, *r*, and p value for each.

    Practice retrieval and exam score: r(____) = _____, p = _____.

    Spacing and exam score: r(____) = _____, p = _____.

    Rereading and exam score: r(____) = _____, p = _____.

29. Compute and interpret the effect sizes (r^2) for each of the above correlations. Are they small, small to medium, medium, and so on?

    Practice retrieval and exam score: _____

    Spacing and exam score: _____

    Rereading and exam score: _____

30. Do the above results support, partially support, or refute Dunlosky et al.'s (2013) studying strategy recommendations?

    a. Support

    b. Partially support

    c. Refute

31. Write an APA-style summary for all three correlations. After you report the correlations, provide a brief summary of what studying recommendations you would make based on these data.

Activity 13.2: Confidence Intervals for Correlations

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Compute and interpret confidence intervals (CIs) for correlations

COMPUTING CIS FOR CORRELATIONS

When working with sample means, you computed confidence intervals to estimate the range of plausible values for population parameters. Specifically, you used CIs to estimate a population's mean (i.e., μ) or the mean difference between two populations (i.e., $\mu_1 - \mu_2$). You can also use CIs to estimate the range of plausible values for the correlation value (i.e., r) of a population.

As with all of the other confidence intervals you have computed, CIs for correlations have two components—namely, a **point estimate** and a **margin of error.** For correlation CIs, the sample's r is the point estimate. As with all other CIs, the size of the margin of error is determined by two things: (1) the level of confidence desired and (2) the expected amount of sampling error (i.e., *SEM*). Again, as with all previous CIs, the margin of error is then added to and subtracted from the point estimate to determine the upper bound and lower bound of plausible r values for the population. The computational formula for the upper and lower bounds of a correlation CI is as follows:

$$\text{Upper boundary} = \text{Point estimate} + \text{Margin of error.}$$

$$\text{Upper boundary} = (z_r) + (z_{CI})(SEM_{corr}).$$

$$\text{Lower boundary} = \text{Point estimate} - \text{Margin of error.}$$

$$\text{Lower boundary} = (z_r) - (z_{CI})(SEM_{corr}).$$

However, there is one important difference between computing correlation CIs and computing the CIs you worked with previously. You may have noticed that the above computational formulas contain z_r as the point estimate and z_{CI} as part of the margin of error. These zs indicate that the formulas use z scores of rs, not rs themselves. Why does the CI for correlations use z scores? The short answer is that the sampling distribution of r is not normally shaped, and this complicates the CI computations. Consequently, there are additional steps when computing CIs for correlations. First, you must convert the point estimate r to a z score. You can do this quite easily with Appendix G. Simply find the sample r in the table, and the z score you need, z_r, is next to it in the table. Then, if you want a 95% CI, z_{CI} is always 1.96. If you want a 99% CI, z_{CI} is always 2.58. The *SEM* for a correlation, SEM_{corr}, is $\frac{1}{\sqrt{N-3}}$. So, the complete upper and lower bound CI formulas for correlations are as follows:

$$\text{Upper boundary} = (z_r) + (z_{CI})\left(\frac{1}{\sqrt{N-3}}\right).$$

$$\text{Lower boundary} = (z_r) - (z_{CI})\left(\frac{1}{\sqrt{N-3}}\right).$$

Then, after you find the upper and lower values using these formulas, you must convert these z score values back into r values using Appendix G. Simply find the z score values in the table, and the r values next to them are the upper and lower boundaries for the CI of r.

In the previous activity, you computed three correlations to test the predictions of Dunlosky et al. (2013). The correlation matrix from that analysis is below. Now, your research team needs to complete its data analysis by computing the 95% CIs for each of the three correlations involving exam score.

Correlations

		Exam Score	Practice_ Retrieval	Spacing_ scale	ReReading_ scale
Exam 2 [Total Pts: 90]	Pearson Correlation	1	.264**	.255**	.023
	Sig. (1-tailed)		.004	.005	.411
	N	103	102	102	101
Practice_Retrieval	Pearson Correlation	.264**	1	.243**	.018
	Sig. (1-tailed)	.004		.006	.429
	N	102	106	106	105
Spacing_scale	Pearson Correlation	.255**	.243**	1	.122
	Sig. (1-tailed)	.005	.006		.108
	N	102	106	106	105
ReReading_scale	Pearson Correlation	.023	.018	.122	1
	Sig. (1-tailed)	.411	.429	.108	
	N	101	105	105	105

**. Correlation is significant at the 0.01 level (1-tailed).

1. Claudia has already started working on the 95% CI for the practice retrieval and exam score correlation. She used Appendix G to convert the sample correlation of $r = .264$ to a $z = .2661$. She then plugged all the values into the upper bound formula and computed it. Then she converted the upper bound z back into an r by using Appendix G. Her work is below.

$$\text{Upper boundary} = (z_r) + (z_{CI})\left(\frac{1}{\sqrt{N-3}}\right) = .2661 + 1.96\left(\frac{1}{\sqrt{102-3}}\right).$$

$$\text{Upper boundary} = .2661 + 1.96\left(\frac{1}{\sqrt{99}}\right) = .2661 + 1.96(.1005) = .2661 + .1970 = .4631.$$

r to z conversion: z of .4631 is equal to an r of .43 (from Appendix G)

So, after using Appendix G to convert the upper bound z back to an r, the upper bound for the correlation in the population is approximately $r = .43$. Now, help Claudia by completing all of the computations and conversions for the lower boundary.

$$\text{Lower boundary} = (z_r) - (z_{CI})\left(\frac{1}{\sqrt{N-3}}\right) =$$

2. You should have found the lower boundary for the z to be .069 and the lower boundary for the r to be .0701. Now compute the 95% CIs for the other correlation between spacing and exam score.

Spacing and exam score:

Computation of CI	*Conversion Back to* r
Upper bound z =	Upper bound r =
Lower bound z =	Lower bound r =

3. Compute the 95% CI for the correlation between rereading and exam scores.

Rereading and Exam score:

Computation of CI	*Conversion Back to* r
Upper bound z =	Upper bound r =
Lower bound z =	Lower bound r =

4. Why is it important to compute confidence intervals?
   a. CIs allow you to determine if you should reject or fail to reject the null hypothesis.
   b. CIs tell you the probability of obtaining a correlation as extreme or more extreme than the one obtained if the null hypothesis is true.
   c. CIs indicate how much of an effect one variable had on the other variable.
   d. CIs provide a range of plausible values for the population correlation coefficient.

5. Suppose that you computed a 95% CI and found that the lower bound was −.90 and the upper bound was .03. You are concerned that this confidence interval is far too wide and want a more precise estimate of the population correlation. Which of the following would make the confidence interval more precise (narrower)? Choose all that apply.
   a. Compute a 99% confidence interval
   b. Increase the sample size
   c. Improve the measurement accuracy of the data collection

Activity 13.3: Spearman's Correlation

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Use SPSS to run a Spearman's correlation
- Summarize the results of a Spearman's correlation in APA format

Prior to entering kindergarten, all children in a school district complete an assessment of their readiness for kindergarten. The test assesses several different components of readiness, including social skills, mathematics, science, social studies, language, and motor development. Scores on all of these components are measured on an interval scale, where higher numbers indicate greater readiness. In addition, parents also report their income category and highest level of education in the household on the following scales:

_____Less than $10,000
_____$10,000 to $19,999
_____$20,000 to $29,999
_____$30,000 to $39,999
_____$40,000 to $49,999
_____$50,000 to $59,999
_____$60,000 to $69,999
_____$70,000 to $79,999
_____$80,000 to $89,999
_____$90,000 to $99,999
_____$100,000 to $149,999
_____$150,000 or more

_____Did not complete high school
_____High school graduate or GED
_____Some college, no degree
_____Associate's degree
_____Bachelor's degree
_____Master's degree
_____Professional degree (JD, MD) or doctorate

1. You wonder if parents' income level and level of education are correlated with scores on the mathematics readiness test. Why should you use a Spearman's correlation rather than a Pearson's correlation to analyze these data?
   a. Mathematics readiness scores are monotonic.
   b. Income and education are measured on ordinal scales.
   c. The variables are not independent of each other.

2. It is important for the teachers to make sure that the children take the test without help from their parents, siblings, or other children in the room. Of course, this is necessary to ensure that the test provides a good assessment of the child's readiness, but it is also important with respect to statistical assumptions. Which assumption is met by making sure the children complete the test without help?
   a. Independence
   b. Monotonic relationship
   c. Appropriate variable measurement

3. Data from 78 children are included in the data file titled "Correlation_Kindergarten_Readiness.sav." Create a scatterplot of the relationship between income and mathematics readiness. To do this, click on the Graphs menu. Choose Legacy Dialogs, and then select Scatter/Dot. Choose Simple Scatter, and click Define. Next, you need to indicate which variable you want on the x-axis and which variable you want on the y-axis. It does not matter which variable goes on which axis. Click OK to create the scatterplot. Is the relationship monotonic?
   a. Yes
   b. No

4. Create a scatterplot of the relationship between education and mathematics readiness. Is the relationship monotonic?
   a. Yes
   b. No

5. Compute Spearman's correlations for these two relationships. Do this by clicking on the Analyze menu. Choose Correlate, and then select Bivariate. Move the variables you want to correlate into the Variables box (in this case, all three variables). Check the box for Spearman's and uncheck the box for Pearson's. Then click OK.

6. Record the correlations below:

   Mathematics readiness and income: r(_____) = _____, p = _____

   Mathematics readiness and education: r(_____) = _____, p = _____

7. How can you determine if the correlations are significant?
   a. They are significant if the p value is less than .05.
   b. They are significant if the p value is greater than .05.

8. Fill in the blanks of the APA-style summary of the results.

   Overall, higher parental education was associated with _____mathematics readiness scores, r(_____) = _____, p = _____. Likewise, higher parental income was associated with _____ mathematics readiness scores, r(_____) = _____, p = _____. Both effect sizes were _____.

Activity 13.4: Introduction to Regression and Prediction

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Compute a regression equation by hand and using SPSS
- Summarize the results of a regression analysis in APA format

INTRODUCTION TO REGRESSION

Correlation coefficients provide a measure of the strength and direction of the relationship between two variables. Through hypothesis testing, you can determine whether the relationship between two variables is statistically significant. If the relationship is significant, you should be able to predict one variable if you know the scores on the other variable. For example, if height and foot length are significantly correlated, you can use foot length to predict someone's height. You cannot use correlation coefficients directly to do this. Instead, you must create a regression equation. The following example illustrates how this is done.

Scenario 1

Anthropometry is the study of the relationships among human body measurements. Data from anthropometric studies are used to design desks, movie seats, cars, and a variety of other products. These same data can also help investigators solve crimes. For example, a police officer may know the length of a suspect's foot from a footprint left at a crime scene. Most people are not accustomed to looking at foot length, so describing the suspect as having a foot 10.5 inches long would probably not help find the suspect. It would be far more useful to tell people to look for a suspect who is of a certain height. Investigators might be able to use the suspect's foot length to predict the suspect's height. The accuracy of this prediction would depend on the size of the correlation between foot length and height. To determine if foot length is a good predictor of height, a researcher collects the following data:

Foot Length (X) (inches)	*Height (Y) (inches)*
8.5	64
10	67
11	71
8	61
9.5	67

1. Compute the Pearson's correlation between height and shoe size. You may use SPSS or do it by hand.

You should have found a correlation coefficient of $r = .98$. Clearly, there was a strong correlation between foot length and height. This strong correlation can help you predict the suspect's height from the suspect's foot length. You will need to use a regression equation specifically designed to predict height from foot length.

2. What is the primary purpose of a regression equation?

The regression equation you will use has this general format:

$$\hat{Y} = bX + a'$$

where $\hat{Y}$ is the variable we are trying to predict (in this case height) and X is the variable we use to make the prediction (in this case foot length). Both b and a are values that change based on the specific values of X and Y, and therefore their exact values must be computed for each regression equation. Specifically, b is the slope of the regression line and is computed by dividing the standard deviation of the variable height (SD_Y) by the standard deviation of foot length (SD_X) and then multiplying the product by the correlation coefficient (r).

3. Compute b.

$$b = r\left(\frac{SD_Y}{SD_X}\right) = _____.$$

You should have found that $b = 3.07$. Next, you will need to compute a, which is the y intercept of the regression line. To compute a, you multiply the mean of foot length (M_X) by b and subtract the result from the mean of height (M_Y).

4. Compute a.

$$a = M_Y - bM_X = _____.$$

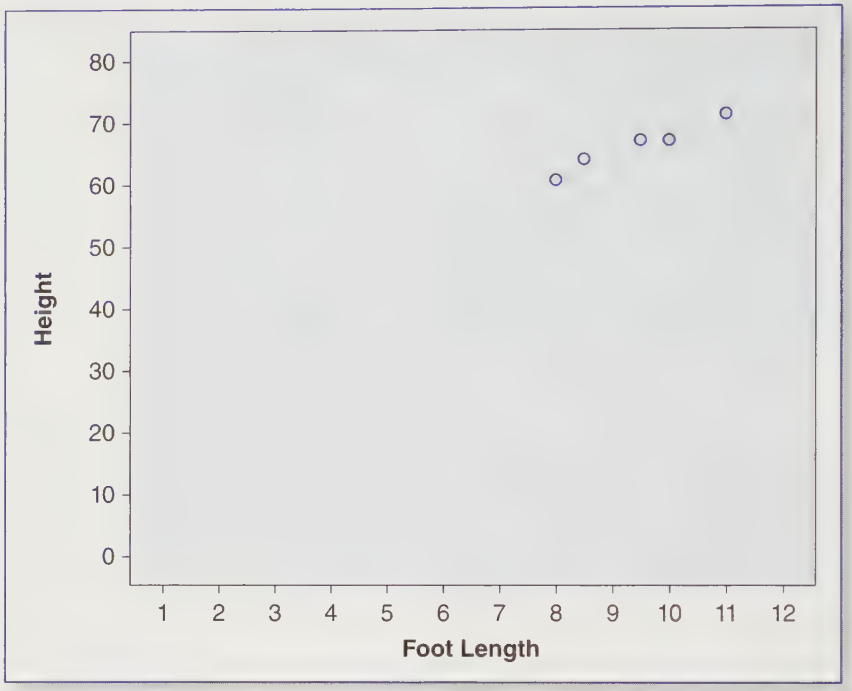
5. You should have found that $a = 37.14$. The general regression equation is displayed below. Use the specific values for b and a that you computed above to create a regression equation that predicts height from foot length. In other words, substitute the specific values of b and a into the general regression equation, and write that specific regression equation below.

$$\hat{Y} = bX + a.$$

6. You can now use this equation to predict a person's height based on his or her foot length. For example, use the equation to predict the height of someone with a foot that is 7 inches long. In other words, when $X = 7$, what does $\hat{Y}$ equal?

7. Use the same equation to predict the height of someone with a foot that is 10.5 inches long. What is the predicted height of the suspect who left the footprint at the crime scene?

8. Below is a scatterplot of the foot length and height data from the first page of this activity. Plot the two predicted data points from Questions 6 and 7 on the scatterplot, and then draw a straight line through those two points.



The line you just drew is called the foot length–height regression line because it was created using the foot length–height regression equation. The regression equation gives you the line that best fits the data. In this case, the line fits the data very well. Overall, the individual data points are very close to the line. This should not be too surprising given that the correlation between height and foot length was .98. In general, the greater the magnitude of the correlation (the closer it is to −1 or +1), the closer the data points will be to the line.

9. You can estimate how well the regression equation predicts Variable *Y* (foot length) by computing r^2 (the coefficient of determination). Compute r^2.

10. Which of the following correlations (i.e., *r* values) will result in the most accurate predictions?

    a. −.84

    b. +.40

    c. 0

SPSS

You can also compute a regression equation using SPSS. First, enter the data into SPSS by creating one column for height and another for foot length. Second, compute the regression equation with the following steps:

- Click on the Analyze menu. Choose Regression, and then select Linear.
- Move foot length into the Independent measures box and height into the Dependent Variables box.
- Click on OK to run the analysis.

The relevant portion of the output is provided below.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	37.140	3.448		10.770	.002
	foot_length	3.070	.365	.980	8.423	.004

a. Dependent Variable: height

(Constant): *a* in the regression equation: the *y* intercept

(foot_length): *b* in the regression equation; the slope of the line

(Beta): the correlation coefficient (*r*)

Scenario 2

When the partial remains of a skeleton are found, investigators often need to know the height of the person to help identify who the person was. Although foot length is a good predictor, it is not always available, and other predictors must be used. A researcher wants to know how accurately the length of the hand can predict height. She collects the following data:

Hand Length (X) (inches)	*Height (Y) (inches)*
5.9	65
7.8	69
6.3	68
6.5	67
6.9	70
6.2	62

11. The correlation between hand length and height was $r = .648$. Compute b.

$$b = r\left(\frac{SD_Y}{SD_X}\right) = _____.$$

12. Compute a.

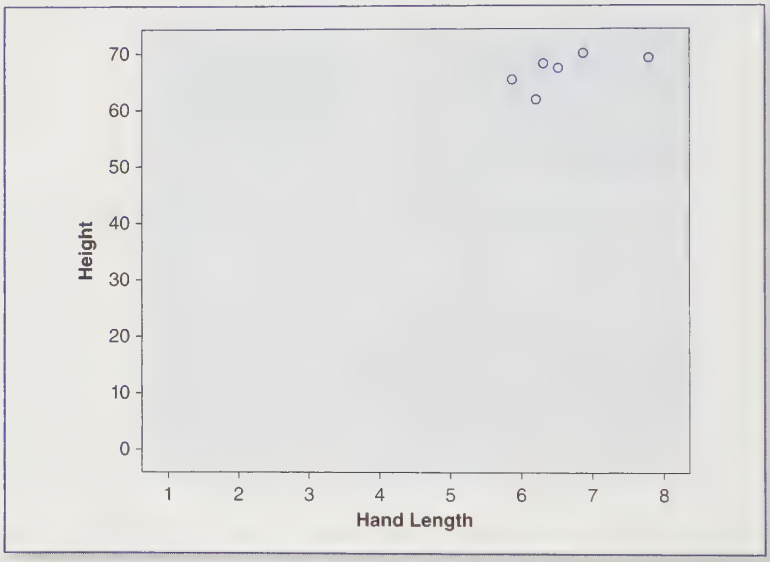
$$a = M_Y - bM_X = _____.$$

13. Write the specific regression equation for predicting height from hand length. In other words, substitute a and b into the general regression equation, and write the specific regression equation.

14. Find the predicted height of someone with a hand length of 5.5 inches.

15. Find the predicted height of someone with a hand length of 7.5 inches.

16. Plot the two predicted data points from Questions 14 and 15 in the scatterplot below, and then draw the regression line.



17. Compute r^2.

18. Explain how you can use the scatterplots for foot length and hand length to determine which variable (hand length or foot length) is the better predictor of height.

19. Explain how you can use r^2 to determine which variable (hand length or foot length) is the better predictor of height.

20. Enter the data from Scenario 2 into SPSS, and run a regression equation to predict height from hand length. The relevant portions of the SPSS regression output follow.

Model Summary

Model	R	R Square	Adjusted *r* Square	Std. Error of the Estimate
1	.648[a]	.419	.274	2.49341

a. Predictors: (Constant), HandLength

R: the correlation coefficient *(r)*

R Square: the correlation coefficient *(r)* squared; the proportion of variability in the criterion (DV) explained by the predictor (IV)

ANOVA[b]

Model		Sum of Squares	*df*	Mean Square	F	Sig.
1	Regression	17.965	1	17.965	2.890	.164[a]
	Residual	24.868	4	6.217		
	Total	42.833	5			

a. Predictors: (Constant), HandLength

b. Dependent Variable: Height

Sig: the *p* value

If $p \leq \alpha$, conclude that the predictor (IV) explains a significant proportion of the variability in the criterion (DV)

If $p > \alpha$, conclude that the predictor (IV) does not explain a significant proportion of the variability in the criterion (DV)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	48.307	10.946		4.413	.012
	HandLength	2.807	1.651	.648	1.700	.164

a. Dependent Variable: Height

(Constant): *a* in the regression equation; the *y* intercept

(HandLength): *b* in the regression equation; the slope of the line

Beta: the correlation coefficient *(r)*

You can use the SPSS output to locate *b* and *a* and thus create the regression equation. The output also gives a way to determine how well the data points fit the regression line (i.e., it gives you an effect size for the correlation you are using to make your prediction). The r^2 value of .419 (in the "Model Summary" output) tells you that 41.9% of the variance in height can be explained by hand length. The value of r^2 should be interpreted using the same guidelines as for n^2: Namely, values close to .1 are small, values close to .09 are medium, and values close to .25 are considered large. The correlation between hand length and

height is a large correlation. To determine whether 49.1% is a statistically significant amount of the variance, you look at the "ANOVA" output. The F value was 2.89, with a p value of .164 (Sig.). If the Sig. value is less than or equal to alpha (.05), you conclude that the predictor (hand length) does explain a statistically significant proportion of the variance in the criterion (height). If the Sig. value is greater than alpha (.05), you conclude that the predictor does not explain a significant proportion of the variance in the criterion.

21. Does hand length explain a statistically significant proportion of the variance in height?

22. What conclusion can you draw from the fact that the r^2 (i.e., the effect size) between hand length and height was large and yet hand length did not predict a significant portion of the variance in height (i.e., hand length was NOT a good predictor of height)?

23. Which of the following is the best summary of these results?

    a. Hand length explained a significant proportion of the variance in height, $r^2 = .42$, $F(1, 4) = 2.89$, $p < .05$.

    b. Hand length did not explain a significant proportion of the variance in height, $r^2 = .42$, $F(1, 4) = 2.89$, $p > .05$, but the sample size was probably too small to draw any conclusions from this study.

The following SPSS output is from a regression analysis using head circumference (measured in millimeters) to predict height (in millimeters).

Model Summary

Model	R	R Square	Adjusted *r* Square	Std. Error of the Estimate
1	.352 [a]	.124	.119	62.314

a. Predictors: (Constant), Head_Cir

ANOVA[b]

Model		Sum of Squares	*Df*	Mean Square	F	Sig.
1	Regression	94998.467	1	94998.467	24.465	.000 [a]
	Residual	671757.510	173	3882.991		
	Total	766755.977	174			

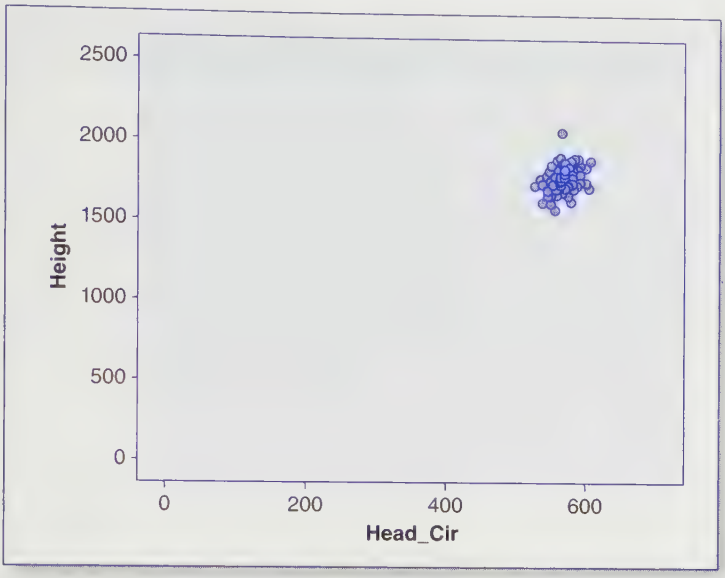
a. Predictors: (Constant), Head_Cir

b. Dependent Variable: Height

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	811.290	190.647		4.255	.000
	Head_Cir	1.660	.336	.352	4.946	.000

a. Dependent Variable: Height



24. Write the regression equation for predicting height from head circumference below.

25. If someone had a head circumference of 570 mm, what would be your prediction of that person's height?

26. What was the value of r^2?

27. Does head circumference explain a significant proportion of the variance in height?

28. Summarize the results in APA style.

29. Head circumference was a significant predictor of height, but hand length was not, despite the fact that r^2 was actually higher for hand length ($r^2 = .419$) than for head circumference ($r^2 = .124$). Why was head circumference a significant predictor but hand length not a significant predictor?

Activity 13.5: Choose the Correct Statistic

Learning Objectives

After reading the chapter and completing the homework and this activity, you should be able to do the following:

- Read a research scenario and determine which statistic should be used

CHOOSE THE CORRECT STATISTIC

Correlation coefficients are used when the research question is about whether there is an association (relationship) between two variables. If both variables are measured on an interval/ratio scale and the relationship is linear, you should choose a Pearson's correlation. If both variables are measured on an interval/ratio scale, but the relationship is monotonic but not linear, you should use a Spearman's correlation. If at least one variable is measured on an ordinal scale, you should use a Spearman's correlation. For these research scenarios, you do not have information about the scatterplot. Instead, you will make your decision between Pearson and Spearman based on whether the variables are interval/ratio or ordinal.

Determine which of the following statistics should be used in each of the research scenarios given below: *z* for sample mean, single-sample *t*, related samples *t*, independent samples *t*, one-way independent samples analysis of variance (ANOVA), two-way independent samples ANOVA, Pearson's correlation, or Spearman's correlation.

1. A researcher wants to know if the relative wealth of a nation is associated with literacy. To determine if these two variables are related, the researcher obtains data from two different sources. The International Monetary Fund (IMF) ranks countries by their gross domestic product (GDP), while the United Nations Development Programme ranks countries by their literacy rates. Which statistic would help the researcher determine if GDP is associated with literacy rates?

2. An English professor is troubled by the quality of his students' papers and believes that they are not putting much effort into their work. He wonders if requiring students to post their papers on the Internet may increase the effort students put into the paper and hence the quality of those papers. To test this, the professor assigns students in two different sections of the class to write a research paper. One class is told that they will turn their final paper in to the professor. The other class is told that they will turn their final paper in by posting it on the Internet. A different professor reads the papers (without knowing if they were posted on the Internet) and rates them for quality on a 10-point scale, with 1 being the worst and 10 being the best. Analysis of the results revealed that the papers that were posted online received higher scores ($M = 6.59$) than the papers that were turned in to the professor ($M = 5.12$). Which statistic could be used to determine if the online papers were significantly better than the papers that were just turned into the professor?

3. Although posting papers on the Internet was effective, it was quite a bit of work for the professor and the students. Thus, the English professor wonders if posting on the Internet is really necessary. It is possible that this increase in quality is the result of students being concerned about people other than the professor reading their papers. If this is the case, the professor

reasons, having other students in class read their papers would have the same effect. For the next paper, the students were placed in one of three groups: The first group was asked to give the paper only to the professor; the second group, to give the paper to another student and the professor; and the third group, to post the paper online. Again, the quality of the papers was judged on a 10-point scale. Which statistic could be used to determine if there was a difference in quality across the three groups?

4. A number of patients tell a physical therapist (PT) that yoga seems to help improve balance. The PT wants to know if yoga may help her elderly patients, who are particularly prone to fractures following falls. To determine if yoga improves balance, the PT assesses balance in each of 29 patients. Each patient completes a series of tasks (e.g., standing on one foot) and then receives a score between 0 and 75, with higher numbers indicating better balance. All patients then participate in 1-hour-long yoga sessions twice a week for 3 months. At the end of the 3 months, balance is assessed once again. What statistic could be used to determine if balance scores were higher after completing the yoga sessions than before?

5. A therapist working in a college counseling center finds that a common problem facing students is loneliness. He wonders if loneliness changes throughout their 4 years in college. He collects data from 20 freshmen, 20 sophomores, 20 juniors, and 20 seniors. Loneliness is measured using a questionnaire that results in scores ranging from 1 to 60, with higher numbers indicating greater loneliness. What statistic could be used to determine if freshmen were lonelier than sophomores, juniors, and seniors?

6. A counselor at a different college hears of these results and wonders if the same would be true of her students. More than half of her students live at home, and she thinks that these students may be less susceptible to loneliness than students who do not live at home. To test, this she collects data on loneliness through a questionnaire from freshmen, sophomores, juniors, and seniors. She also asks them whether or not they live at home. What statistic could be used to determine if the relationship between year in school and loneliness is the same for students who live at home and for students who live at school?

7. A researcher conducts a study to determine if sugar consumption is associated with depression. The researcher records per capita sugar consumption for 30 different countries and also records the prevalence of depression in those countries. Prevalence rates were recorded as the estimated percentage of the population with depression. What statistic should be used to determine if, as sugar consumption rises, so does the prevalence of depression?

8. The study about sugar consumption and depression is picked up by the media, with many suggesting that sugar consumption is the cause of depression. Of course, the previous study did not allow the researcher to determine if one variable caused the other. To answer this question, another researcher designed a carefully controlled study with 53 participants. At the beginning of the study, the researcher measured the depression levels for all 53 participants. For 1 month following the initial assessment, the participants were instructed to consume at least 400 grams of white sugar each day. At the end of the month, depression scores were again recorded. What statistic should be used to determine if depression scores were higher after 1 month of a high-sugar diet?

9. A statistics instructor wants to know if hard work really pays off in her course, and so at the end of the semester, the instructor gives students a questionnaire assessing how much effort they have put into the course. The effort questionnaire yields one value between 0 and 100, with higher numbers indicating more effort. She also records the students' final grade in the course (A, B, C, D, or F). What statistic should be used to determine if effort is associated with course grades?

10. Research has shown that exercise leads to increased high-density lipoprotein (HDL) cholesterol levels (HDL is the good cholesterol); however, exactly how much exercise is needed is not known. Thus, a researcher designs a study in which participants with low levels of HDL are randomly assigned to exercise for 0 minutes a day, 30 minutes, 60 minutes, or 90 minutes. After 3 months of these exercise regimens, the HDL cholesterol levels are measured. Women and men often react differently to treatments, and so the researcher also investigated the impact of gender in the study. What statistic could be used to determine if the effect of exercise on cholesterol is the same for males and females?

11. To test the efficacy of new cancer treatments, patients with cancer must be recruited to participate in clinical trials. A doctor notices that some of his colleagues seem to be very likely to enroll patients in these trials while others seem to be less likely to enroll their patients in clinical trials. She wonders if this discrepancy is a function of the specialty of the physician. Specifically, she wants to know if surgeons are more likely than nonsurgeons to enroll their patients in clinical trials. To test this, she records whether each of 54 doctors are surgeons or not and then also records the percentage of each doctor's cancer patients who enrolled in clinical trials. What statistic could she use to determine if the percentage of doctors who enroll their patients in clinical trials is the same for surgeons and nonsurgeons?

12. An industrial/organizational psychologist is interested in understanding the effect of spending time on the Internet at work doing things unrelated to work. She wonders if time spent off the job is harmful to performance on the job or if it provides a needed break and actually improves performance. To assess the relationship between Internet use and performance, she sends 98 employees an anonymous survey. On the survey, they are asked to indicate how many minutes they spend each day on the Internet doing things that are not related to work. She also asks them to report their score on their annual performance review. Scores on the performance review range from 1 to 10, with higher scores indicating better performance. What statistic could be used to determine whether there is a statistically significant relationship between time on the Internet and performance scores?

13. Every year, students in a particular school district take a standardized test that is designed to assess reading and writing skills. Last year, the average score on this test was 49 (slightly below average). The principal of this district attempts to improve these scores by trying a new program with a sample of 50 third graders. This new program rewards kids for reading throughout the school year and the summer. She hopes that this extra reading will improve the scores on the test. What statistic could she use to determine if reading scores were higher in the group that participated in the new program?

Chapter 13 Practice Test

SCENARIO 1

A student in a statistics class thinks that people who like to run are weird. To test this hypothesis, she gives a sample of students two questionnaires to complete. One assesses how much they like to run, with higher numbers indicating a more favorable attitude toward running. The other questionnaire assesses weirdness, with higher numbers indicating more weirdness. Here are the data she obtained. Given her expectation that those who like to run are also weird, she chooses to do a one-tailed test. She also chooses an alpha of .05.

Running Rating	*Weirdness Rating*
5	7
9	9
4	6
5	3
2	4
1	6
4	4
7	6
6	5
4	8
4	5

1. Match each of the following statements to the statistical assumption to which it is relevant.

   ___e___ Data independence

   ___a___ Appropriate measurement of variables

   ___c___ Normality

   ___b___ Proper trend in bivariate scatterplot

   a. Both variables are measured on an interval/ratio scale of measurement.
   b. One variable is measure on a nominal scale and the other is on an interval/ratio scale.
   c. Each variable has a normal shape in the population.
   d. One of the two variables has a normal shape in the population.
   e. Individuals' responses were not influenced by others' responses.
   f. An individual's response to the first question is unrelated to that individual's response to the second question.

g. There is a fan appearance to the scatterplot with greater variability for some values of X than for other values of X.

h. There is a linear trend to the scatterplot, with no extreme scores.

2. Create a scatterplot, and determine if a Pearson's or a Spearman's correlation is appropriate.

a. The trend in the scatterplot is sufficiently linear, so a Pearson's correlation can be used.

b. The trend in the scatterplot is not sufficiently linear, so a Pearson's correlation should not be used.

3. Which of the following is the correct symbolic representation of the null hypothesis for this one-tailed test?

a. H_0: $\rho > 0$.

b. H_0: $\rho < 0$.

c. H_0: $\rho \geq 0$.

d. H_0: $\rho \leq 0$.

4. Which of the following is the best verbal representation of the null hypothesis for this one-tailed test?

a. How much people like to run is positively associated with how weird they are.

b. How much people like to run is negatively associated with how weird they are.

c. How much people like to run is not associated with how weird they are, or it is negatively associated with how weird they are.

5. Which of the following is the correct symbolic representation of the research hypothesis for this test?

a. H_1: $\rho > 0$.

b. H_1: $\rho < 0$.

c. H_1: $\rho \geq 0$.

d. H_1: $\rho \leq 0$.

6. Which of the following is the best verbal representation of the research hypothesis for this test?

a. How much people like to run is positively associated with how weird they are.

b. How much people like to run is negatively associated with how weird they are.

c. How much people like to run is not associated with how weird they are.

7. What is the *df* for this correlation (i.e., $df = N - 2$)?

a. 22

b. 18

c. 11

d. 9

8. What is the critical value for this one-tailed Pearson's correlation ($\alpha = .05$)?

a. .476

b. .497

c. .521

d. .549

9. Compute Pearson's r.
   a. .189
   b. .428
   c. .521
   d. .549

10. What is the effect size for this study (i.e., r^2)?
   a. .036
   b. .183
   c. .271
   d. .301

11. Is the effect size small, small to medium, medium, medium to large, or large?
   a. Small
   b. Small to medium
   c. Medium
   d. Medium to large
   e. Large

12. The effect size of this study indicates that
   a. 18% of the variability in scores for liking to run can be explained by the variability in weirdness scores.
   b. 18% of the scores for liking to run are unexplainable.
   c. 18% of the scores for liking to run are explainable.

13. Which of the following is the best summary of this study's results?
   a. How much people like to run and their weirdness scores are significantly correlated, $r(9) = .43$, $p < .05$.
   b. How much people like to run and their weirdness scores are not significantly correlated, $r(9) = .43$, $p > .05$. However, the medium to large effect size suggests that the study's sample size was too small.
   c. How much people like to run and their weirdness scores are significantly correlated, $r(9) = .52$, $p > .05$.
   d. How much people like to run and their weirdness scores are not significantly correlated, $r(9) = .52$, $p < .05$.

14. When do you use a Spearman's correlation rather than a Pearson' correlation (choose all that apply)?
   a. When both variables are measured on an interval/ratio scale
   b. When one variable is ordinal and the other is interval/ratio
   c. When both variables are ordinal
   d. When the scatterplot is monotonic rather than linear
   e. When the scatterplot is linear rather than monotonic

15. Suppose that a track coach wants to try to identify middle school students who might be good runners by giving students the weirdness questionnaire. Because there was a positive correlation, the coach reasons that students who are weird would be good runners. Why is it not a good idea to use the weirdness ratings to predict running ability?
    a. Because the correlation coefficient was not significant
    b. Because running is a better predictor of weirdness than weirdness is of running

16. After the coach realizes that weirdness is not a good predictor of running ability, she decides to collect data to determine if grip strength is a good predictor of running speed. To assess the relationship between grip strength and running speed, she measures the grip strength of 78 students. The average grip strength for the students was $M = 60.85$, $SD = 17.03$, with higher numbers indicating greater grip strength. The same students then run a quarter of a mile. The average running speed was $M = 86.73$ seconds, $SD = 26.17$ seconds. The correlation between the two measures was $r(76) = -.29$, $p < .05$. Compute b for the regression equation.
    a. .19
    b. −.19
    c. .45
    d. −.45

17. Compute a for the regression equation.
    a. 114.11
    b. −114.11
    c. 99.88
    d. −99.88

18. Which of the following is the correct regression equation?
    a. $\hat{Y} = -99.88X + 99.88$
    b. $\hat{Y} = -.45X + 114.11$
    c. $\hat{Y} = 114.11 + 99.88X$

19. Use the regression equation to predict the running speed of a person with a grip strength score of 50.
    a. 88.79
    b. 66.59
    c. 91.61
    d. 78.92

20. How much of the variability in running speed is explained by grip strength?
    a. 8%
    b. 16%
    c. 29%

REFERENCE

Dunlosky, J., Rawson, K.A., Marsh, E.J., Nathan, M.J., & Willingham, D.T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4–58.

CHAPTER 14

Goodness-of-Fit and Independence Chi-Square Statistics

LEARNING OBJECTIVES

After reading this chapter, you should be able to perform the following:

- Describe the difference between a goodness-of-fit chi-square and a chi-square test of independence, and identify when to use each
- Describe the logic of the chi-square statistic
- Write the null and research hypotheses for goodness-of-fit chi-square and a chi-square test of independence
- Compute the chi-square statistic and determine if the null hypothesis should be rejected
- Summarize the results of a chi-square

OVERVIEW OF CHI-SQUARE

Most of the statistics you have learned in this course require you to compute one or more means before the results can be interpreted. However, there are many interesting research scenarios in which the data are nominal (e.g., *political affiliation:* Democrat, Republican, Independent) or ordinal (e.g., *highest level of education completed:* high school, some college, college, some postgraduate, graduate), and therefore, it is impossible to compute a mean. For example, suppose you wanted to know if voters' affiliation with political parties is significantly different today than it was 2 years ago. You could obtain a sample of voters and ask them with which political party they most often affiliate: Democrat, Republican, or Independent. Example data are presented in Table 14.1.

In this case, it would be impossible to compute a mean response because the response options are nominal categories. The data are *frequency counts* for each political party (i.e., the number of people who said Democrat, Republican, or Independent). You can use a chi-square statistic when the data are frequency counts for different categories.

Table 14.1 Number of People Identifying With Each Political Affiliation

Political Affiliation		
Democrat	Republican	Independent
22	22	12

Reading Question

1. When the data that a study produces is in the form of _____ the chi-square statistic is appropriate.
   a. means
   b. standard deviations
   c. interval/ratio data
   d. frequency counts for different categories

There are two different types of chi-square statistics, and they are each used in different research situations. A *goodness-of-fit chi-square* is used to analyze the relative frequencies of different categories within a single variable. The political affiliation study described above would be analyzed using a goodness-of-fit chi-square because the researchers want to compare the relative frequencies of Democrats, Republicans, and Independents, which are all different categories of the variable political affiliation. A *chi-square test of independence* is used to determine if the relative frequencies within the categories of one variable are associated with the relative frequencies within the categories of a second variable. For example, suppose you wanted to know if people's political affiliation (e.g., Democrat, Republican, or Independent) was associated with their opinion of a Democratic president's policies (e.g., approve or disapprove). To answer this question, you could obtain a sample of voters and ask them their political affiliation (i.e., Democrat, Republican, or Independent) and their opinion of the president's policies (i.e., approve or disapprove). All the possible combinations of answers create six categories, and the voters' actual answers determine the frequency counts in each of the six categories. Example data are presented in Table 14.2.

Table 14.2 Number of People Identifying With Each Political Affiliation by Approval of the President's Policies

		Political Affiliation		
		Democrat	Republican	Independent
Policy opinion	Approve	15	2	6
	Disapprove	7	20	6

In this case, there were 15 Democrats who approved of the president's policies and 7 Democrats who disapproved of his policies. The chi-square for independence will determine if the two variables of political affiliation and policy opinion are independent (i.e., if there is no association between the two variables).

Reading Question

2. The _____ analyzes the frequency counts of categories within a single variable.
   a. goodness-of-fit chi-square
   b. chi-square for independence

Reading Question

3. The _____ analyzes the frequency counts of categories across two different variables.
   a. goodness-of-fit chi-square
   b. chi-square for independence

LOGIC OF THE CHI-SQUARE TEST

Both chi-square tests use the same logic. In fact, they both use the following formula:

$$\text{Chi-square} = \chi^2 = \sum \frac{(OF - EF)^2}{EF}.$$

There are only two terms in the chi-square formula, the observed frequency (i.e., OF) and the expected frequency (i.e., EF), but each cell of the chi-square has its own observed and expected frequencies. The observed frequencies are the actual frequency counts collected from the sample; they are the actual data. The expected frequencies are defined by the null hypothesis and the study's sample size. The expected frequencies are the exact frequency counts expected *if the null hypothesis is true,* given the study's sample size. Studies based on more data will have higher expected frequencies than those based on less data. A very important assumption of both types of chi-squares is that if an expected frequency is less than 5, a chi-square should not be performed until after the sample size is increased.

Reading Question

4. If an expected frequency is less than 5, a chi-square should not be performed because the sample size is too
   a. large.
   b. small.

Reading Question

5. The expected frequencies are the frequencies that are expected if the _____ hypothesis is true.
   a. null
   b. research

The numerator of the chi-square statistic computes the difference between the observed frequency (i.e., the actual data) and the expected frequency (i.e., what the data should be if the null is true). If the null is true and *there were no sampling error,* the observed frequency and the expected frequency would be identical and the numerator would be zero. Of course, some sampling error usually does occur. The possibility of sampling error means that "small" differences between the observed and expected frequencies are probably due to sampling error, but "large" differences indicate that the null hypothesis is probably false.

The denominator of the chi-square formula, which contains only the expected frequency, helps determine how big the numerator difference must be to be considered "large." By dividing the expected frequency into the numerator difference, a ratio of *relative difference* is created. For example, if the computed

difference in the numerator was 12 (i.e., 18 – 6 = 12) and the expected frequency was 6, the resulting ratio would be 24 (i.e., $12^2/6 = 24$). In this case, a difference of 12 is a pretty large *relative* difference. In contrast, if the computed difference in the numerator was 12 (i.e., 180 – 168 = 12) but the expected frequency was 168, the resulting ratio would be .86 (i.e., $12^2/168 = .86$). In this case, a difference of 12 is a relatively small difference. When the expected frequency is large, it takes a larger difference between observed and expected frequencies to be "large."

The Σ symbol in the equation indicates that you must create a relative difference ratio for each category in the study and then sum all of them to create a final obtained chi-square value. If the null hypothesis is true, the final obtained chi-square value is likely to be close to zero. If the null hypothesis is false, the obtained chi-square value is likely to be far from zero. As with other statistics, if the obtained chi-square is greater than the critical value, you should reject the null hypothesis.

Reading Question

6. The numerator of the chi-square computes the difference between the observed and expected frequencies. If the null is true, the difference is likely to be close to
   a. 0.
   b. 1.
   c. 6.
   d. .25.

Reading Question

7. The denominator of the chi-square helps determine
   a. whether a difference between the observed and expected frequencies is "large," relative to the size of the expected frequency.
   b. the critical value of the chi-square statistic.

Reading Question

8. How many pairs of observed and expected frequencies will you compute when using a chi-square statistic?
   a. 2
   b. 6
   c. 0
   d. As many pairs as there are categories in the study

Reading Question

9. If the summed relative differences (called the obtained chi-square value) is greater than the critical value, you should
   a. reject the null.
   b. fail to reject the null.

COMPARING THE GOODNESS-OF-FIT CHI-SQUARE AND THE CHI-SQUARE FOR INDEPENDENCE

Both types of chi-squares (a) analyze frequency counts, (b) test a null hypothesis, (c) use the same formula, and (d) compare an obtained chi-square value with a critical value to determine if the null hypothesis should be rejected. However, as discussed earlier, the goodness of fit analyzes a single variable while the chi-square for independence analyzes the association between two variables. We will now work through a complete example of each type of chi-square starting with the goodness of fit.

GOODNESS-OF-FIT CHI-SQUARE EXAMPLE

We will now analyze the data introduced at the beginning of the chapter concerning the relative frequencies of voters' affiliations with different political parties. Suppose you collected the data in Table 14.3 from a sample of voters from your county.

You want to know if these frequency counts of political affiliation are different from the frequencies reported 2 years ago, specifically 50% Democrats, 30% Republicans, and 20% Independents.

Table 14.3 Number of People Identifying With Each Political Affiliation

Political Affiliation		
Democrat	Republican	Independent
22	22	12

Step 1: Examine Statistical Assumptions

You collected your data carefully, ensuring that the responses from each individual were uninfluenced by the responses of other participants, thereby satisfying the *data independence assumption*. The data are frequency counts of category membership, satisfying the *appropriate measurement of variables assumption*. The final assumption is that all of the *expected* frequency values must be larger than 5. If any of the expected frequencies are less than 5, you should collect more data before you run your chi-square analysis. In this case, there are a total of 56 people in the study, and the smallest proportion expected in a cell is 20%. Thus, the smallest expected frequency is 56(.20) = 11.2. Because this value is greater than 5, you can proceed with the chi-square test.

Step 2: State the Null and Research Hypotheses

In this scenario, the null hypothesis is that the proportions of people affiliating with each party will be the same as they were 2 years ago. Specifically, *if the null hypothesis is true,* you would expect your sample to be 50% Democrats, 30% Republicans, and 20% Independents. The research hypothesis is that the proportions of people affiliating with each party will be different from what they were 2 years ago.

Reading Question

10. The null hypothesis for the goodness-of-fit chi-square always states that the frequency counts in each category
    a. are the same.
    b. are similar to the expected frequency counts for each category.

Step 3: Compute the *df* and Define the Critical Region

The chi-square is different from other statistics in that the *df* that determines the critical value is *not* based on sample size. Rather, in a chi-square, the *df* is based on the *number of categories* in the study. Specifically, the *df* for a goodness-of-fit chi-square is

$$df = \text{Categories} - 1.$$

In this case, there are three categories, and so the $df = 2$. When you look up the critical value for chi-square in Appendix G, you find that when the $df = 2$ and the $\alpha = .05$, the critical value is 5.99.

Therefore, if the computed chi-square value is equal to or greater than 5.99, you should reject the null hypothesis.

Reading Question

11. The *df* value for a chi-square is based on
    a. the number of participants.
    b. the number of categories being counted.

Step 4: Compute the Test Statistic (Goodness-of-Fit Chi-Square)

4a. Determine the Expected Frequencies Defined by the Null Hypothesis

The expected frequencies are created by converting the percentages (or proportions) stated in the null hypothesis into the exact frequencies expected, given the study's sample size. In this case you collected a total of 56 responses (i.e., 22 + 22 + 12 = 56), and thus, the expected frequencies for each category of political affiliation would be as follows:

Percent Predicted by Null	*Expected f*
50% Democrats: 50% of 56 = .50(56) =	28
30% Republicans: 30% of 56 = .30(56) =	16.8
20% Independents: 20% of 56 = .20(56) =	11.2

In this case, all of the expected frequencies are larger than 5, so you can compute the chi-square test statistic.

Reading Question

12. The expected frequencies in a goodness-of-fit chi-square
    a. can be different in every cell.
    b. will always be the same in every cell.

Table 14.4 Observed Frequencies (OF) and Expected Frequencies (EF) for Each Political Affiliation

Political Affiliation		
Democrat	Republican	Independent
OF = 22 EF = 28	OF = 22 EF = 16.8	OF = 12 EF = 11.2

4b. Compute the Obtained Value

The observed and expected frequencies are presented in Table 14.4.

The following obtained chi-square value is calculated by creating a relative difference ratio for each category in the study and then summing them.

$$\text{Chi-square} = \chi^2 = \sum \frac{(OF - EF)^2}{EF} = \frac{(22-28)^2}{28} + \frac{(22-16.8)^2}{16.8} + \frac{(12-11.2)^2}{11.2}$$

$$= 1.285 + 1.610 + 0.057$$

$$= 2.952.$$

The obtained chi-square value of 2.952 is not greater than the critical chi-square value of 5.99. Therefore, you should not reject the null hypothesis. The differences between the observed and expected frequency counts were small enough that they could have resulted from sampling error.

Reading Question

13. If the obtained chi-square is less than the critical value, you should
    a. reject the null hypothesis.
    b. fail to reject the null hypothesis.

After computing the test statistic, the next step is usually computing an effect size. However, we will skip this step for goodness-of-fit chi-square tests. Unlike all other effect size statistics presented in this text, the measure of effect size for a goodness-of-fit chi-square is not a measure of association between two variables, and therefore, its interpretation is quite different from all other measures of effect size presented in this text. Refer to Cohen (1988) if you are interested in computing effect sizes for a goodness-of-fit chi-square.

Step 5: Interpret the Results

The results of this chi-square test are summarized as follows:

The political affiliations with the Democratic, Republican, and Independent parties reported by the sample of voters from this county are not significantly different from those reported 2 years ago, $\chi^2(2, N = 56) = 2.952$, $p > .05$.

CHI-SQUARE FOR INDEPENDENCE

We will now analyze the data introduced earlier concerning the association between voters' political affiliation and their opinion of the president's policies.

Step 1: Examine Statistical Assumptions

The chi-square for independence has the same assumptions as the goodness-of-fit chi-square—specifically, the data must be frequency counts (*appropriate variable measurement*), individuals' responses must not be influenced by other responses (*data independence*), and all *expected* frequency values must be larger than 5.

Step 2: State the Null and Research Hypotheses

The chi-square for independence helps determine whether two variables measured on nominal or ordinal scales of measurement are associated. In this case, the chi-square tests whether one's approval or disapproval of the president's policies is associated with one's political affiliation. Similar to the null hypotheses for the correlation, this chi-square's null hypothesis states that the two variables in the study are not associated (i.e., that the two variables are independent of each other). Specifically, the null hypothesis in this case would be "Policy opinion and political affiliation are not associated." The research hypothesis states that the two variables in the study are associated.

Reading Question

14. The null hypothesis for the chi-square for independence always states that the two variables being analyzed are
    a. associated.
    b. not associated.

Step 3: Compute the *df* and Define the Critical Region

The *df* for a chi-square for independence is determined by the following formula:

$$df = (Columns - 1) * (Rows - 1),$$ where Columns = number of categories in the columns and Rows = number of categories in the rows.

In this case, Columns = 3 and Rows = 2. The *df* is (3 − 1) * (2 − 1) = (2) * (1) = 2. The critical value for a chi-square when *df* = 2 and the $\alpha = .05$ is 5.99. If the obtained chi-square is greater than 5.99, the null hypothesis should be rejected.

Reading Question

15. As with the goodness-of-fit chi-square, the *df* for the chi-square for independence is based on the
    a. number of participants in the study.
    b. number of categories being counted.

Step 4: Compute the Test Statistic (Chi-Square for Independence)

4a. Determine the Expected Frequencies Defined by the Null Hypothesis

Each of the six categories (i.e., cells) created by the possible combinations of responses to political affiliation and policy opinion will have its own expected frequency. The expected frequencies are displayed in *italics* in the lower right corner of each cell. Each cell's expected frequency is based on the total frequency count of the row (i.e., RT) and column (i.e., CT) of which it is a part. The total frequency counts for each row and column are presented in Table 14.5.

The formula used to compute the expected frequencies for each cell in the study is as follows:

$$EF = \frac{RT * CT}{N},$$ where RT refers to the frequency in a given row and CT refers to the frequency in a given column.

Table 14.5 Observed Frequencies (OF) and Expected Frequencies (EF) for Each Political Affiliation by Approval of the President's Policies

		Political Affiliation			
		Democrat	Republican	Independent	Column Total
Policy opinion	Approve	15 *9.103*	2 *9.103*	7 *5.793*	24
	Disapprove	7 *12.897*	20 *12.897*	7 *8.207*	34
	Row total	22	22	14	$N = 58$

For example, for the first cell in the design (Democrats/Approve), the RT is 24 and CT is 22, and N is the total number of people ($N = 58$).

Each expected frequencies are computed as follows:

The EF for the Democrat/Approve cell is $$EF = \frac{(24)(22)}{58} = 9.103.$$

The EF for the Republican/Approve cell is $$EF = \frac{(24)(22)}{58} = 9.103.$$

The EF for the Independent/Approve cell is $$EF = \frac{(24)(14)}{58} = 5.793.$$

The EF for the Democrat/Disapprove cell is $$EF = \frac{(34)(22)}{58} = 12.897.$$

The EF for Republican/Disapprove cell is $$EF = \frac{(34)(22)}{58} = 12.897.$$

The EF for the Independent/Disapprove cell is $$EF = \frac{(34)(14)}{58} = 8.207.$$

Reading Question

16. An expected frequency is computed separately for each cell.
    a. True
    b. False

4b. Compute the Obtained Value

The obtained chi-square value is calculated by finding the relative difference between the observed and expected frequencies for each cell in the study and then summing them. The obtained chi-square is computed as follows:

$$\chi^2 = \sum \frac{(OF - EF)^2}{EF} = \frac{(15 - 9.103)^2}{9.103} + \frac{(2 - 9.103)^2}{9.103} + \frac{(7 - 5.793)^2}{5.793} + \frac{(7 - 12.897)^2}{12.897} + \frac{(20 - 12.897)^2}{12.897} + \frac{(7 - 8.207)^2}{8.207}$$

$$= 3.820 + 5.542 + 0.251 + 2.696 + 3.912 + 0.178$$

$$= 16.399.$$

The obtained chi-square value of 16.40 is greater than the critical chi-square value of 5.99. Therefore, the null hypothesis should be rejected. The differences between the observed and expected frequency counts were large enough that they were unlikely to have been created by sampling error; instead, it is likely that voters' opinions of the president's policies and their political affiliation are associated.

Step 5: Compute the Effect Size and Interpret It as Small, Medium, or Large

When computing the effect size for all of the other statistics discussed in this text, we provided a single statistic (i.e., d for t tests, η_p^2 for ANOVAs, and r^2 for correlations). However, when computing the effect size for a chi-square for independence, the statistic you should use depends on the size of the chi-square test you performed. If the chi-square was a 2 × 2 (i.e., both variables had exactly two categories), the phi (ϕ) coefficient is the correct measure of the study's effect size. If either variable has more than three categories, Cramer's ϕ is the correct measure of effect size. The formulas and effect size guidelines for phi and Cramer's ϕ are presented in Table 14.6. You should note that the effect size guidelines change for Cramer's ϕ as the df of the chi-square test increases.

In this case, Cramer's ϕ is the appropriate measure of effect size because one of the variables in the original chi-square has three categories. Cramer's ϕ is computed as follows:

$$\phi = \sqrt{\frac{\chi^2}{N(df*)}} = \sqrt{\frac{16.399}{58(1)}} = 0.532.$$

The largest effect size guidelines are appropriate because $df^* = (CT - 1) = 2 - 1 = 1$. Therefore, the size of the association between voters' political affiliation and their opinion of the president's policies is large.

Reading Question

17. When both variables in a chi-square analysis have just two levels, _____ is the appropriate measure of effect size.
    a. the phi coefficient
    b. Cramer's phi

Reading Question

18. When at least one variable in a chi-square analysis has more than two levels, _____ is the appropriate measure of effect size.
    a. the phi coefficient
    b. Cramer's phi

Table 14.6 Effect Sizes for Chi-Square and Guidelines for Interpretation

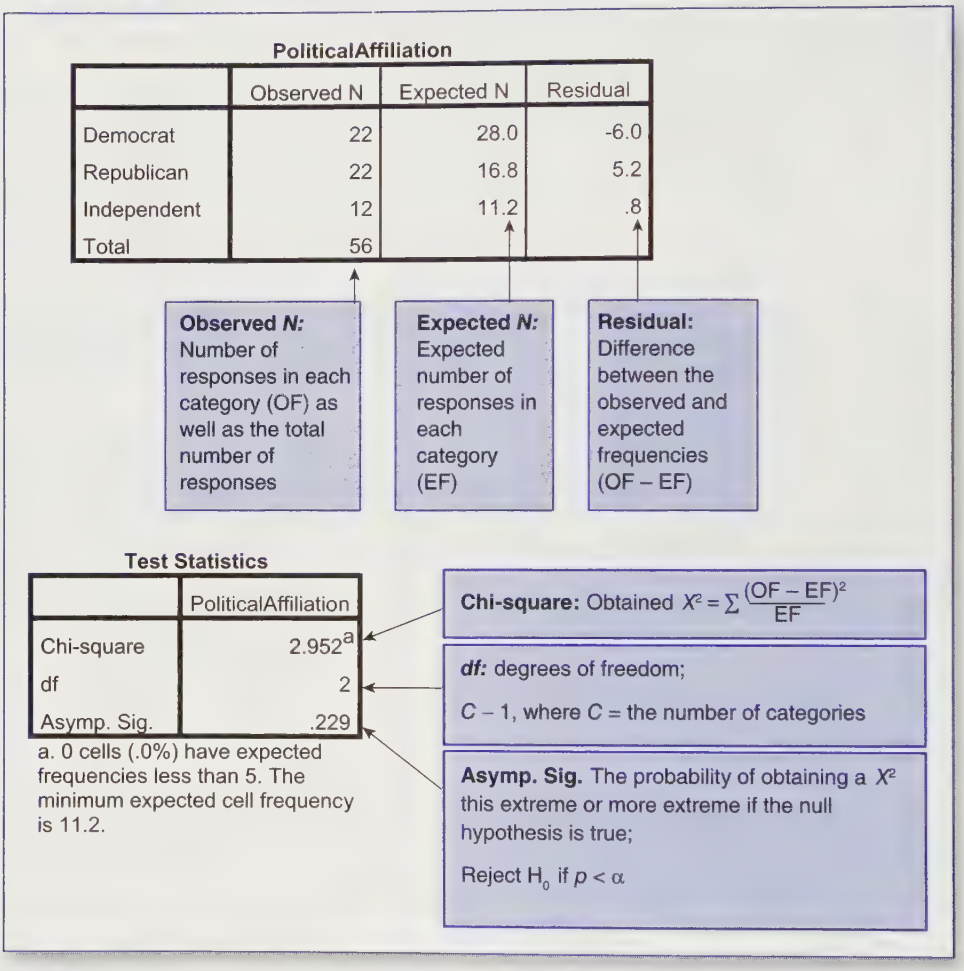
Statistic	*When It Is Used*	*Formula*	*Effect Size Guidelines*
Phi coefficient (ϕ)	Both variables have two categories.	$\phi = \sqrt{\frac{\chi^2}{n}}$	.1 = small .3 = medium .5 = large
Cramer's phi (ϕ)	At least one variable has three or more categories.	$\phi' = \sqrt{\frac{\chi^2}{n(df*)}}$ df^* = (Columns − 1) or (Rows − 1), whichever is smaller	$df^* = 1$ .1 = small .3 = medium .5 = large $df^* = 2$ .07 = small .21 = medium .35 = large $df^* = 3$ .06 = small .17 = medium .29 = large

Step 6: Interpret the Results

In this study, the null hypothesis was rejected, meaning that the association between policy opinion and political affiliation is statistically significant. After rejecting a chi-square's null hypothesis, you must look at the differences between the observed and expected frequencies in each cell. Looking at these values reveals that four of the six cells had large differences, and four had only small differences. These four cells with the large differences were those in the Democrat and Republican columns. It is clear from these values that more Democrats approved of the president's policies than was expected by chance and that fewer Democrats disapproved of the president's policies than was expected by chance. The exact opposite pattern was found for Republicans. The following passage is an example of how the results of this study might be reported.

The voters' opinions of the president's policies were associated with the voters' political affiliations, $\chi^2(2, N = 58) = 16.40$, $p < .05$, $\phi = .53$. More Democrats and fewer Republicans approved of the president's policies than would be expected by chance. More Republicans and fewer Democrats disapproved of the policies than would be expected by chance. The approval and disapproval of Independents were very close to what was expected by chance.

Figure 14.3 SPSS Screenshot of Observed and Expected Frequencies for Chi-Square Test of Goodness of Fit



Chi-Square for Independence

We will now analyze the political affiliation and policy opinion data using SPSS to perform a chi-square for independence. The data are shown in Table 14.8.

Data File

The data file for the chi-square for independence will have two columns. One will have the values for political affiliation (i.e., 1 = Democrat, 2 = Republican, 3 = Independent) in a single column, and one will have the values for policy opinion (i.e., 1 = approve, 2 = disapprove). You will need to enter the data values for each categorical variable on the Variable View screen as you did in the previous SPSS example. The actual data file will have 15 rows in which the value in the political affiliation

column is a 1 and the value in the policy opinion column is also a 1. There would be two rows with political affiliation = 2 and policy opinion = 1, then seven rows with political affiliation = 3 and policy opinion = 1. This coding system would be continued for the disapprove row in the above table. After all of the data are entered into the SPSS Data Editor, your data file will look similar to the screenshot in Figure 14.4.

Table 14.8 Data Used for SPSS Example for Chi-Square Test of Independence

		Political Affiliation			*Total*
		Democrat	Republican	Independent	
Policy opinion	Approve	15	2	7	24
	Disapprove	7	20	7	34
	Total	22	22	14	$n = 58$

Figure 14.4 SPSS Screenshot of Data Entry for Chi-Square Test of Independence

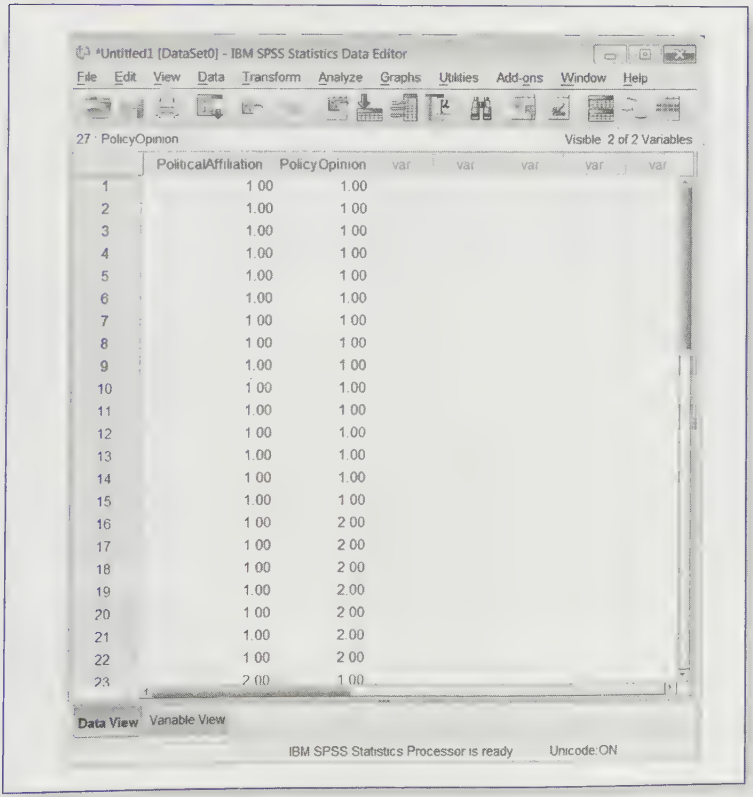


Figure 14.5 SPSS Screenshot of Output With Total Number of Responses; Output of Observed Frequencies, Row Totals, and Column Totals; and Chi-Square Test of Independence

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Political_Affiliation * Policy_Opinion	58	100.0%	0	.0%	58	100.0%

Total *N*: The total number of responses

Political_Affiliation * Policy_Opinion Crosstabulation

			Policy_Opinion		
			Approve	Disapprove	Total
Political_Affiliation	Democrat	Count	15	7	22
		Expected Count	9.1	12.9	22.0
	Republican	Count	2	20	22
		Expected Count	9.1	12.9	22.0
	Independent	Count	7	7	14
		Expected Count	5.8	8.2	14.0
Total		Count	24	34	58
		Expected Count	24.0	34.0	58.0

Total: The row totals

Count: The observed frequency for each category (OF)

Expected Count: The expected frequency for each category (EF)

Total: The column totals

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	16.400[a]	2	.000
Likelihood Ratio	18.339	2	.000
Linear-by-Linear Association	2.553	1	.110
N of Valid Cases	58		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.79.

Asymp. Sig. The probability of obtaining a χ^2 this extreme or more extreme if the null hypothesis is true

Reject H_0 if $p < \alpha$

df: degrees of freedom;

(Columns – 1) * (Rows – 1)

Chi-square: Obtained χ^2

$$\sum \frac{(OF - EF)^2}{EF}$$

SPSS Analysis

To perform the chi-square:

- Click on the Analyze menu. Choose Descriptive Statistics and select Crosstabs.
- Move the variable labeled "PoliticalAffiliation" to the Rows box and the variable labeled "PolicyOpinion" to the Column(s) box.
- Click on the Statistics button and check the chi-square box and click Continue.
- Click the Cells button and make sure that the observed and expected boxes are both checked.
- Click Continue; then click OK.

SPSS Output

Reading Question

21. Should the null hypothesis be rejected for this analysis?
    a. Yes
    b. No

OVERVIEW OF THE ACTIVITIES

In Activity 14.1, you will work through examples of both types of chi-squares. In Activity 14.2, you will choose what statistical procedure is appropriate for various research situations.

Activity 14.1: Goodness-of-Fit Chi-Square and Chi-Square for Independence

Learning Objectives

After reading the chapter and completing this activity, you should be able to do the following:

- Determine when to use a goodness-of-fit chi-square or a chi-square for independence
- Determine the expected frequencies for chi-square
- Compute a goodness-of-fit chi-square and a chi-square for independence
- Summarize the results of a chi-square analysis

WHEN TO USE THE GOODNESS-OF-FIT VERSUS THE CHI-SQUARE FOR INDEPENDENCE

1. The chi-square statistic is used to analyze nominal or ordinal data. If you are analyzing a single variable, you would use a _____ (Choose one: goodness-of-fit chi-square/chi-square for independence)

   If you are analyzing the relationship between two nominal or ordinal variables, you would use a _____ (Choose one: goodness-of-fit chi-square/chi-square for independence)

In this text, we have devoted most of our time to learning statistics that we can use when one or more of our variables are interval or ratio. The reason for this is that interval/ratio data are more precise than nominal or ordinal data, and as a result, most researchers design their studies to collect data that are interval or ratio. However, there are several very interesting research areas in which many if not most of the data are nominal or ordinal in nature. One prominent example of a research area in which many of the variables are nominal is in political science and public opinion polling.

WHEN EXPECTED FREQUENCIES ARE TOO LOW

2. The chi-square statistic is appropriate for all of the research scenarios described in the following questions. You should remember, however, that if one or more computed expected frequencies for any chi-square analysis are less than _____ (choose one: .01, .05, 5, 10), you should not perform a chi-square test. If an expected frequency is less than this number, you should collect more data until all of the expected frequencies are greater than this number.

FOUR EXAMPLES

Between August 27 and 30, 2010, researchers called registered voters and asked them several questions and recorded their answers. One of the questions they asked was as follows: "Now, thinking back on some of the major pieces of legislation Congress has passed in the past 2 years, would you say you approve or disapprove of the health care overhaul?" Of the 1,021 people who were asked this question, 970 responded by saying "approve" or "disapprove." The rest had "no opinion" and were excluded from the analysis. The observed frequency counts for each response options are provided below. Compute a chi-square to determine if the sample's responses were significantly different from a 50%/50% approve/disapprove split on the health care overhaul.

Approve	*Disapprove*
398	572

3. Compute the expected frequency for each cell. Write the expected frequency for each cell into the above cells.

4. Are all of the expected frequencies sufficiently large to proceed with the chi-square?
   a. Yes
   b. No
5. Determine the critical value for this chi-square analysis.
6. Compute the obtained chi-square value.

7. Determine if you should reject or fail to reject the null hypothesis.
8. Which measure of effect size should be used for this goodness-of-fit chi-square?
   a. The phi coefficient, ϕ
   b. Cramer's phi, ϕ'
   c. Neither; we will not compute and effect size for goodness-of-fit chi-square
9. Choose the best summary of this chi-square analysis and fill in the blanks.
   a. Significantly more people approve of the health care overhaul than would be expected by chance, χ^2(___, N = ____) = _____, $p < .05$.
   b. Significantly more people disapprove of the health care overhaul than would be expected by chance, χ^2(___, N = ____) = _____, $p < .05$.
   c. The proportions of people who approve and disapprove of the health care overhaul are not significantly different from those that would be expected by chance, χ^2(____, N = _____) = _____, $p < .05$.

Another question that the researchers asked registered voters was, "Now, thinking back on some of the major pieces of legislation Congress has passed in the past 2 years, would you say you approve or disapprove of increased government regulation of banks and major financial institutions?" The observed frequency counts for the response categories of "approve," "disapprove," and "no opinion" are below. Test the null hypothesis that the responses will be 50% approve, 50% disapprove, and 0% no opinion.

Approve	Disapprove	No Opinion
620	373	28

10. Compute the expected frequency for each cell. Write the expected frequency for each cell into the above cells.

11. Are all of the expected frequencies sufficiently large to proceed with the chi-square?
    a. Yes
    b. No

12. There are two ways you could deal with the expected frequency of the "no opinion" cell being too low. You could change your hypothesis so that the expected frequency would be higher, or you could simply drop the respondents from the analysis. Which do you think would be the better choice in this case? Explain your reasoning.
    a. Drop the "no opinion" category.
    b. Change the research hypothesis.

13. Now drop the "no opinion" respondents and recompute the expected frequencies for the approve and disapprove cells. Write them in the respective cells.

14. Determine the critical value for this chi-square analysis.

15. Compute the obtained chi-square value.

16. Determine if you should reject or fail to reject the null hypothesis.

17. Fill in the blanks for the APA-style summary of the results of this study.

    Significantly _____ people approve of the increased regulation on financial institutions than would be expected by chance, χ^2(____) = _____, $p <$ _____.

18. In the study you just completed, the "no opinion" respondents were excluded because the expected frequency in that cell would have been 0. The only reason it would have been 0 was because the hypothesis was that 0% of the respondents would not have any opinion. It would have been okay to include the "no opinion" respondents if the hypothesis being tested had predicted some value greater than 0 for that cell. Suppose that researchers always expect about 5% of every population to have no opinion on any issue. If that were the case for the previous analysis, what would have been the expected frequencies for each of the approve, disapprove, and no opinion cells?

Respondents were also asked what Congress should do about tax cuts. They were given three options: (1) keep them in place for all taxpayers, (2) keep them in place for those making less than $250K a year but end them for those making more than $250K a year, or (3) do nothing and allow the tax cuts to expire for all taxpayers. The observed frequencies for each of these categories are shown below. Test the hypothesis that respondents are *equally split across these three categories*.

Keep Tax Cuts for All	*Keep Tax Cuts for Only Those Making Less Than $250K*	*Do Nothing and Allow Tax Cuts to Expire for All Taxpayers*
378	449	153

19. Compute the expected frequency for each cell. Write the expected frequency for each cell into the above cells.

20. Are all of the expected frequencies sufficiently large to proceed with the chi-square?
    a. Yes
    b. No

21. Determine the critical value for this chi-square analysis.

22. Compute the obtained chi-square value.

23. Determine if you should reject or fail to reject the null hypothesis.

24. The following APA-style summary has three errors. Find these three errors and correct them. Two of the errors are in the wording and one is a numerical error.

    Fewer people favor extending the tax cuts to everyone or extending them to only those making less than $250K than would be expected by chance and more people favor allowing the tax cuts to expire than would be expected by chance, $\chi^2(1) = 146.35$, $p < .05$. The greatest discrepancy from the expected value was in the middle option in which more people favored the tax cuts for only those making less than $250K.

Finally, it seems likely that a respondent's political identification as Republican, Democrat, or Independent will be associated with one's opinion on the tax cut issue. To test the hypothesis that political identification and opinion on the tax cuts are associated, compute a chi-square on the data in the following table.

	Keep Tax Cuts for All	*Keep Tax Cuts for Only Those Making Less Than $250K*	*Do Nothing and Allow Tax Cuts to Expire for All Taxpayers*
Republican	204	121	42
Independent	111	116	43
Democrat	65	215	68

25. Compute the expected frequency for each cell. Write the expected frequency for each cell into the above cells.
26. Are all of the expected frequencies sufficiently large to proceed with the chi-square?
    a. Yes
    b. No
27. Determine the critical value for this chi-square analysis.
28. Compute the obtained chi-square value.

29. Determine if you should reject or fail to reject the null hypothesis.
30. Which measure of effect size should be used for this test for independence chi-square?
    a. The phi coefficient, ϕ
    b. Cramer's phi, ϕ'
31. Compute the appropriate measure of effect size for this study.
32. Determine which set of effect size guidelines should be used and then determine if the effect is small, medium, or large.
33. Correct three errors in the APA-style summary below. One error is numerical, two are in the wording.

    Political identification and opinion on extending the Bush tax cut were not associated, $\chi^2(4) = 9.49$, $p < .05$. Republicans were much more likely to favor keeping the tax cuts for all taxpayers. Independents were slightly less likely to favor keeping the tax cuts for only those making less than $250K a year. Democrats were much more likely to favor extending the tax cuts for only those making less than $250K a year.

Activity 14.2: Choose the Correct Statistic

Learning Objectives

After reading the chapter, completing the homework and this activity, you should be able to do the following:

- Read a research scenario and determine which statistic should be used

CHOOSE THE CORRECT STATISTIC

Determine which statistic should be used in each of the following research scenarios: *z* for sample mean, single-sample *t*, related samples *t*, independent samples *t*, one-way independent samples ANOVA, two-way independent samples ANOVA, Pearson's correlation, Spearman's correlation, chi-square goodness of fit, or chi-square test of independence.

a. *z* for a sample mean
b. Single-sample *t*
c. Independent measures *t*
d. Repeated/related measures *t*
e. Independent measures ANOVA
f. Two-factor ANOVA
g. Pearson's correlation
h. Spearman's correlation
i. Chi-square goodness of fit
j. Chi-square test of independence

1. Is a male politician's height associated with his popularity with voters? A researcher randomly selected 20 male politicians and looked up their height on the Internet. The researcher then looked up each of their popularity ratings on the Internet (ratings were given as a percentage approval rating that ranged from 0%–100%).

2. Although the Internet provides access to a wealth of information, finding that information is not always easy. However, a number of skills can greatly speed up the process. For example, "ctrl-f" can be used to find a particular word once you are on a web page. Unfortunately, a recent study revealed that 90% of people did not know this shortcut. A student wants to know if this is also true of college students. The student obtains a sample of 73 students and asks them what hitting the "ctrl-f" key combination does when you are on a web page. Each person's response was coded as either correct or incorrect. Fifty-two students answered the question correctly while 21 answered the question incorrectly. What statistic should be used to determine if the proportion of college students who know the "ctrl-f" shortcut is different from the proportion of people in the general population who know the shortcut?

3. A neurologist has 20 patients, all of whom suffered severe head trauma between 3 and 6 months ago. He wants to know if their ability to recall words is different from the general population. He gave all 20 patients a standardized memory test. The general population's mean score on this standardized test is 75 with a standard deviation of 10. His patients scored a mean of 60. What statistic should be used to determine if the recall scores of his patients are significantly different from the recall scores in the general population?

4. Are college students who attend religiously affiliated schools more religious than the general population? A national survey reports that on a 10-point religiosity scale (with 10 being the most religious rating), the average rating for Americans is 7 with a standard deviation of 3.2. A sample of 157 students who all attended religiously affiliated colleges had a mean religiosity rating of 7.75. Is the sample of students significantly more religious than the general American population?

5. Several researchers have found that the attitudes students have about learning are associated with their grades in college. Students who view classes as an opportunity to learn new things tend to perform better in college than do students who view classes as obstacles that they must tolerate

25. Compute the expected frequency for each cell. Write the expected frequency for each cell into the above cells.
26. Are all of the expected frequencies sufficiently large to proceed with the chi-square?
    a. Yes
    b. No
27. Determine the critical value for this chi-square analysis.
28. Compute the obtained chi-square value.

29. Determine if you should reject or fail to reject the null hypothesis.
30. Which measure of effect size should be used for this test for independence chi-square?
    a. The phi coefficient, ϕ
    b. Cramer's phi, ϕ'
31. Compute the appropriate measure of effect size for this study.
32. Determine which set of effect size guidelines should be used and then determine if the effect is small, medium, or large.
33. Correct three errors in the APA-style summary below. One error is numerical, two are in the wording.

    Political identification and opinion on extending the Bush tax cut were not associated, $\chi^2(4) = 9.49$, $p < .05$. Republicans were much more likely to favor keeping the tax cuts for all taxpayers. Independents were slightly less likely to favor keeping the tax cuts for only those making less than $250K a year. Democrats were much more likely to favor extending the tax cuts for only those making less than $250K a year.

Activity 14.2: Choose the Correct Statistic

Learning Objectives

After reading the chapter, completing the homework and this activity, you should be able to do the following:

- Read a research scenario and determine which statistic should be used

CHOOSE THE CORRECT STATISTIC

Determine which statistic should be used in each of the following research scenarios: z for sample mean, single-sample t, related samples t, independent samples t, one-way independent samples ANOVA, two-way independent samples ANOVA, Pearson's correlation, Spearman's correlation, chi-square goodness of fit, or chi-square test of independence.

a. z for a sample mean
b. Single-sample t
c. Independent measures t
d. Repeated/related measures t
e. Independent measures ANOVA
f. Two-factor ANOVA
g. Pearson's correlation
h. Spearman's correlation
i. Chi-square goodness of fit
j. Chi-square test of independence

1. Is a male politician's height associated with his popularity with voters? A researcher randomly selected 20 male politicians and looked up their height on the Internet. The researcher then looked up each of their popularity ratings on the Internet (ratings were given as a percentage approval rating that ranged from 0%–100%).

2. Although the Internet provides access to a wealth of information, finding that information is not always easy. However, a number of skills can greatly speed up the process. For example, "ctrl-f" can be used to find a particular word once you are on a web page. Unfortunately, a recent study revealed that 90% of people did not know this shortcut. A student wants to know if this is also true of college students. The student obtains a sample of 73 students and asks them what hitting the "ctrl-f" key combination does when you are on a web page. Each person's response was coded as either correct or incorrect. Fifty-two students answered the question correctly while 21 answered the question incorrectly. What statistic should be used to determine if the proportion of college students who know the "ctrl-f" shortcut is different from the proportion of people in the general population who know the shortcut?

3. A neurologist has 20 patients, all of whom suffered severe head trauma between 3 and 6 months ago. He wants to know if their ability to recall words is different from the general population. He gave all 20 patients a standardized memory test. The general population's mean score on this standardized test is 75 with a standard deviation of 10. His patients scored a mean of 60. What statistic should be used to determine if the recall scores of his patients are significantly different from the recall scores in the general population?

4. Are college students who attend religiously affiliated schools more religious than the general population? A national survey reports that on a 10-point religiosity scale (with 10 being the most religious rating), the average rating for Americans is 7 with a standard deviation of 3.2. A sample of 157 students who all attended religiously affiliated colleges had a mean religiosity rating of 7.75. Is the sample of students significantly more religious than the general American population?

5. Several researchers have found that the attitudes students have about learning are associated with their grades in college. Students who view classes as an opportunity to learn new things tend to perform better in college than do students who view classes as obstacles that they must tolerate

before they get to their careers. One of the studies classified students into one of these two groups and then compared the mean college GPAs of the two groups. What statistic should be used to determine if the GPAs of the two groups are significantly different?

6. A teacher gives a multiple-choice test with 100 questions. Every question on the test has exactly one correct answer and three incorrect answers. The teacher wants to know how well his 37 students did compared with chance performance (i.e., compared with 25 correct answers). What statistic should be used to determine if his students did better than chance?

7. What is the association between years of formal education people completed and their income? A group of students used government survey data to answer this question. The years of education variable was measured as "through junior high," "through high school," "some college," "completed college," "some graduate school," and "completed graduate school." For the income variable, they classified each family into one of seven different income categories ranging from "below the poverty line" to "more than $1 million dollars a year."

8. Do the feeding practices at a certain zoo help make its gorillas healthy? At a certain zoo, the gorillas must "hunt and forage" for their food rather than finding it placed in the same location of their habitat every day. Is this practice beneficial to the gorillas? To address this question, a zoo worker hid the gorillas' food in a different location every day. After 3 years, the zoo veterinarian gave every gorilla a physical and used a standardized rating system to give each gorilla a rating for health that ranged from 1 = *very unhealthy* to 10 = *very healthy*. The minimal acceptable rating on this scale that is established by the National Zoological Society is 7. Are the gorillas at this zoo significantly healthier than the minimum national standard for gorilla health?

9. In an attempt to study attitude change, a group of students asked participants their opinion of a local sandwich shop by having them rate the shop's sandwiches on a scale from 1 (*terrible*) to 10 (*great*). The participants then had to write a persuasive speech in which they were to make the strongest argument they could to convince others that they should eat at the sandwich shop. A week after writing the persuasive speech, the participants were again asked to rate the shop's sandwiches on the same 1-to-10 scale. What statistic should be used to determine if writing a persuasive speech leads to higher ratings than prior to writing the speech?

10. A researcher studying problem solving presented "the candle problem" to two different groups. To solve the problem, participants had to discover a way to attach a burning candle to a wall. Both groups were given a candle, a small box of wooden matches, thumb tacks, and paper clips. The only difference between the two conditions was the starting position of the matches. For half of the participants, the matches started in the box, and for half, the matches were presented outside of the box. The researcher marked what condition each participant was in (i.e., filled-box condition or empty-box condition) and whether each participant solved the problem or did not solve the problem. What statistic should be used to determine if the box-full and box-empty conditions differ in terms of the proportion of participants who successfully solved the problem?

11. A researcher taught one group of people to use the method of loci and then asked them to use it to study and then recall a list of 100 words. Another group of people were given the same amount of time to study the same list of 100 words and were also asked to recall them. Did the method of loci group recall more words than the control group?

12. The testing effect illustrates that trying to retrieve information that you have learned (i.e., testing yourself over what you have learned) increases one's memory for that material on future tests. One of the studies that illustrates the testing effect had half of the participants study material for

four 5-minute sessions before taking a final test while the other participants studied the same material for one 5-minute session and took three sample tests before taking the final test. In addition, half of the participants in each group took the final test after a delay of several minutes while the other half took the final test after a delay of 1 week. The dependent variable in this study was the number of correct responses on the final test. (Incidentally, none of the questions on the final test appeared on any of the sample tests.) What statistic should be used to determine if practice tests result in higher scores than studying without practice tests? Did the testing effect differ based on the length of the time delay before the final test?

13. Do speed reading classes really increase reading speed without a decline in comprehension? A researcher randomly assigned readers to one of two groups. One group completed a nationally advertised speed reading course. The other group received no speed reading training. Half of the people in each group were given as much time as they needed to read a book, and then they took test on the book. The other half of each group did not read the book but took the test on the book to function as two control conditions.

14. A group of students in a criminology course obtained government and police records to analyze the relationship between a family's income and the number of times that family had been burglarized. They classified each family into one of seven different income categories ranging from "below the poverty line" to "more than $1 million dollars a year." The number of times each family was burglarized was obtained from police records. What statistic should be used to determine if the number of times a family was burglarized is associated with their income level?

15. A group of undergraduate students performed a variation of a classic psychology experiment for a class project. In a series of trials, a participant had to say which of two lines was longer. In every trial, one of the lines was substantially longer than the other, so it was very obvious which line was longer. However, before the participant made the judgment on each trial, four confederates (i.e., people working with the experimenter who were pretending to be research participants) made their judgments. For several trials, all of the confederates made the correct judgment. On the critical trial, however, all four confederates intentionally chose the wrong line. The critical question was whether or not the one true participant would conform by also picking the wrong line or if the participant would go against the group and pick the line that was obviously longer. The undergraduate students predicted that only 25% of the participants would resist the social pressure to conform by choosing the correct line and 75% of the sample would conform by choosing the line that was obviously shorter. What statistic should be used to test their hypothesis?

16. Postpartum depression is quite common among new mothers, and recent research has shown that a significant proportion of men also experience depression with the addition of a child to the family. A researcher wonders if the level of depression experienced by the mother is associated with the level of depression experienced by the father. To test this, the researcher gives both the mother and the father the Inventory to Diagnose Depression (IDD). Scores on this inventory range from 0 to 88, with higher scores indicating greater levels of depression. What statistic should be used to determine if there is a relationship between the IDD scores of mothers and fathers?

17. The researcher discovers that IDD scores of the mother and father are significantly associated. However, she also wants to know if men and women have significantly different levels of depression. Thus, she runs a different analysis to compare the IDD scores of the mother and father. When setting up the data file, she is careful to match each mother's IDD scores with her partner's IDD score. What statistic should be used to determine if mothers were significantly more depressed than fathers?

18. A clinician wonders if the socioeconomic status (SES) of a family is associated with levels of depression after having a new child. The clinician obtains a sample of 60 new mothers. Twenty have a low SES, 20 are middle class, and 20 have a high SES. Each mother completes the IDD. What statistic should be used to determine if mothers with low SES have higher levels of depression than mothers with a higher SES?

19. Does working outside of the home affect postpartum depression? To investigate this possibility, a researcher obtains a sample of 75 women: 25 who worked full time outside of the home, 25 who worked part time outside of the home, and 25 who stayed home. IDD scores were collected 6 months after the baby was born.

20. Postpartum depression affects men as well, and so another researcher decides to replicate the study above using both mothers and fathers. This time, 75 new mothers and 75 new fathers were recruited: within each gender, 25 who worked full time outside of the home, 25 who worked part time outside of the home, and 25 who stayed home. What statistic should be used to determine if IDD scores depend on gender and work status?

Chapter 14 Practice Test

GOODNESS-OF-FIT PROBLEM

For a project in a statistics class, you must collect data and then analyze them using a chi-square statistic. You decide to collect data on how easily people could go 1 week without using various software/social media products. Specifically, you ask 60 of your fellow students which of the following three programs they would be most willing to not use for 1 week: (1) e-mail, (2) Facebook, or (3) Twitter. The obtained frequencies for each of these categories are shown below. Your *null hypothesis* is that each of the three categories will have the same frequencies ($N = 60$). Use $\alpha = .05$.

E-mail	*Facebook*	*Twitter*
11	9	40

1. Match each of the following statements to the statistical assumption to which it is relevant.

   ____ Data independence

   ____ Appropriate measurement of variables

   ____ Frequencies are sufficiently high

   a. The data are frequency counts.
   b. One variable is measure on a nominal scale, and the other is on an interval/ratio scale.
   c. Each variable has a normal shape in the population.
   d. One of the two variables has a normal shape in the population.
   e. Individuals' responses were not influenced by others' responses.
   f. Individuals' responses were influenced by others' responses.
   g. The *expected* frequencies are all greater than 5.
   h. The *obtained* frequencies are all greater than 5.

2. What is the *df* for this study?
   a. 1
   b. 2
   c. 3
   d. 4

3. What is the critical value for this study?
   a. 3.841
   b. 5.991
   c. 6.635
   d. 9.210

4. What is the expected frequency for e-mail?
   a. 20
   b. 40
   c. 60
   d. 5
5. What is the obtained χ^2 value for this study?
   a. 4.05
   b. 30.1
   c. 6.05
   d. 20
6. Should the null hypothesis for this study be rejected?
   a. Yes
   b. No
7. Which of the following is the best summary of this study's results?
   a. Students in the study said that they were much more likely to give up using Twitter for 1 week than e-mail or Facebook, $\chi^2(2, N = 60) = 5.99, p < .05$.
   b. Students in the study said that they were much more likely to give up using Twitter for 1 week than e-mail or Facebook, $\chi^2(2, N = 60) = 30.1, p < .05$.
   c. Students in the study said that they were much more likely to give up using Twitter for 1 week than e-mail or Facebook, $\chi^2(3, N = 60) = 7.81, p < .05$.
   d. Students in the study were equally likely to say that they would give up e-mail, Facebook, or Twitter, $\chi^2(3, N = 60) = 7.81, p > .05$.

INDEPENDENCE PROBLEM

Another student in your class collects data for her chi-square project by recording students' gender and asking students whether or not they were vegetarians. Her study had 30 males and 30 females. She wanted to know if the variables of gender and type of diet (i.e., vegetarian or not vegetarian) were associated. Her observed frequencies are shown in the following table. Use $\alpha = .05$.

	Vegetarian	*Not Vegetarian*
Male	6	24
Female	14	16

8. The assumptions for the chi-square for independence are _____ the assumptions for the goodness-of-fit chi-square.
   a. the same as
   b. different from

9. What is the *df* for this study?
   a. 1
   b. 2
   c. 3
   d. 4

10. What is the critical value for this study?
   a. 3.841
   b. 6.635
   c. 5.991
   d. 9.210

11. What is the expected frequency for males who are vegetarian?
   a. 10
   b. 20
   c. 30
   d. 40
   e. 60

12. What is the expected frequency for females who are not vegetarian?
   a. 10
   b. 20
   c. 30
   d. 40
   e. 60

13. What is the obtained χ^2 value for this study?
   a. .28
   b. .8
   c. 4.8
   d. 5.3

14. Should the null hypothesis of this study be rejected?
   a. Yes
   b. No

15. Compute the effect size.
   a. .28
   b. .13
   c. 4.8
   d. 60

16. Which of the following is the best summary of this study's results?

    a. In the students participating in the study, there was an association between diet (i.e., vegetarian or not vegetarian) and gender, $\chi^2(1, N = 60) = 4.8, p > .05$. More females were vegetarian than was expected by chance, and fewer males were vegetarian than was expected by chance.

    b. In the students participating in the study, there was an association between diet (i.e., vegetarian or not vegetarian) and gender, $\chi^2(1, N = 60) = 4.8, p < .05$.

    c. In the students participating in the study, there was no association between diet (i.e., vegetarian or not vegetarian) and gender, $\chi^2(1, N = 60) = .8, p < .05$.

    d. In the students participating in the study, there was no association between diet (i.e., vegetarian or not vegetarian) and gender, $\chi^2(1, N = 60) = 5.3, p > .05$.

REFERENCE

Cohen, J. (1998). *Statistical power analyses for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Appendix A

Unit Normal Table (z Table)

z Score	Body	*Tail*
0.00	0.5000	*0.5000*
0.01	0.5040	*0.4960*
0.02	0.5080	*0.4920*
0.03	0.5120	*0.4880*
0.04	0.5160	*0.4840*
0.05	0.5199	*0.4801*
0.06	0.5239	*0.4761*
0.07	0.5279	*0.4721*
0.08	0.5319	*0.4681*
0.09	0.5359	*0.4641*
0.10	0.5398	*0.4602*
0.11	0.5438	*0.4562*
0.12	0.5478	*0.4522*
0.13	0.5517	*0.4483*
0.14	0.5557	*0.4443*
0.15	0.5596	*0.4404*
0.16	0.5636	*0.4364*
0.17	0.5675	*0.4325*
0.18	0.5714	*0.4286*
0.19	0.5753	*0.4247*
0.20	0.5793	*0.4207*
0.21	0.5832	*0.4168*
0.22	0.5871	*0.4129*
0.23	0.5910	*0.4090*
0.24	0.5948	*0.4052*
0.25	0.5987	*0.4013*
0.26	0.6026	*0.3974*
0.27	0.6064	*0.3936*
0.28	0.6103	*0.3897*
0.29	0.6141	*0.3859*
0.30	0.6179	*0.3821*
0.31	0.6217	*0.3783*
0.32	0.6255	*0.3745*
0.33	0.6293	*0.3707*
0.34	0.6331	*0.3669*
0.35	0.6368	*0.3632*
0.36	0.6406	*0.3594*
0.37	0.6443	*0.3557*
0.38	0.6480	*0.3520*
0.39	0.6517	*0.3483*
0.40	0.6554	*0.3446*
0.41	0.6591	*0.3409*
0.42	0.6628	*0.3372*
0.43	0.6664	*0.3336*
0.44	0.6700	*0.3300*
0.45	0.6736	*0.3264*
0.46	0.6772	*0.3228*
0.47	0.6808	*0.3192*
0.48	0.6844	*0.3156*
0.49	0.6879	*0.3121*
0.50	0.6915	*0.3085*
0.51	0.6950	*0.3050*
0.52	0.6985	*0.3015*
0.53	0.7019	*0.2981*
0.54	0.7054	*0.2946*
0.55	0.7088	*0.2912*
0.56	0.7123	*0.2877*
0.57	0.7157	*0.2843*
0.58	0.7190	*0.2810*
0.59	0.7224	*0.2776*
0.60	0.7257	*0.2743*
0.61	0.7291	*0.2709*
0.62	0.7324	*0.2676*
0.63	0.7357	*0.2643*
0.64	0.7389	*0.2611*
0.65	0.7422	*0.2578*
0.66	0.7454	*0.2546*
0.67	0.7486	*0.2514*
0.68	0.7517	*0.2483*
0.69	0.7549	*0.2451*
0.70	0.7580	*0.2420*
0.71	0.7611	*0.2389*
0.72	0.7642	*0.2358*
0.73	0.7673	*0.2327*
0.74	0.7704	*0.2296*
0.75	0.7734	*0.2266*
0.76	0.7764	*0.2236*
0.77	0.7794	*0.2206*
0.78	0.7823	*0.2177*
0.79	0.7852	*0.2148*
0.80	0.7881	*0.2119*
0.81	0.7910	*0.2090*
0.82	0.7939	*0.2061*
0.83	0.7967	*0.2033*

(Continued)

(Continued)

z Score	Body	Tail
0.84	0.7995	0.2005
0.85	0.8023	0.1977
0.86	0.8051	0.1949
0.87	0.8078	0.1922
0.88	0.8106	0.1894
0.89	0.8133	0.1867
0.90	0.8159	0.1841
0.91	0.8186	0.1814
0.92	0.8212	0.1788
0.93	0.8238	0.1762
0.94	0.8264	0.1736
0.95	0.8289	0.1711
0.96	0.8315	0.1685
0.97	0.8340	0.1660
0.98	0.8365	0.1635
0.99	0.8389	0.1611
1.00	0.8413	0.1587
1.01	0.8438	0.1562
1.02	0.8461	0.1539
1.03	0.8485	0.1515
1.04	0.8508	0.1492
1.05	0.8531	0.1469
1.06	0.8554	0.1446
1.07	0.8577	0.1423
1.08	0.8599	0.1401
1.09	0.8621	0.1379

z Score	Body	Tail
1.10	0.8643	0.1357
1.11	0.8665	0.1335
1.12	0.8686	0.1314
1.13	0.8708	0.1292
1.14	0.8729	0.1271
1.15	0.8749	0.1251
1.16	0.8770	0.1230
1.17	0.8790	0.1210
1.18	0.8810	0.1190
1.19	0.8830	0.1170
1.20	0.8849	0.1151
1.21	0.8869	0.1131
1.22	0.8888	0.1112
1.23	0.8907	0.1093
1.24	0.8925	0.1075
1.25	0.8944	0.1056
1.26	0.8962	0.1038
1.27	0.8980	0.1020
1.28	0.8997	0.1003
1.29	0.9015	0.0985
1.30	0.9032	0.0968
1.31	0.9049	0.0951
1.32	0.9066	0.0934
1.33	0.9082	0.0918
1.34	0.9099	0.0901
1.35	0.9115	0.0885

z Score	Body	Tail
1.36	0.9131	0.0869
1.37	0.9147	0.0853
1.38	0.9162	0.0838
1.39	0.9177	0.0823
1.40	0.9192	0.0808
1.41	0.9207	0.0793
1.42	0.9222	0.0778
1.43	0.9236	0.0764
1.44	0.9251	0.0749
1.45	0.9265	0.0735
1.46	0.9279	0.0721
1.47	0.9292	0.0708
1.48	0.9306	0.0694
1.49	0.9319	0.0681
1.50	0.9332	0.0668
1.51	0.9345	0.0655
1.52	0.9357	0.0643
1.53	0.9370	0.0630
1.54	0.9382	0.0618
1.55	0.9394	0.0606
1.56	0.9406	0.0594
1.57	0.9418	0.0582
1.58	0.9429	0.0571
1.59	0.9441	0.0559
1.60	0.9452	0.0548
1.61	0.9463	0.0537

z Score	Body	Tail
1.62	0.9474	0.0526
1.63	0.9484	0.0516
1.64	0.9495	0.0505
1.65	0.9505	0.0495
1.66	0.9515	0.0485
1.67	0.9525	0.0475
1.68	0.9535	0.0465
1.69	0.9545	0.0455
1.70	0.9554	0.0446
1.71	0.9564	0.0436
1.72	0.9573	0.0427
1.73	0.9582	0.0418
1.74	0.9591	0.0409
1.75	0.9599	0.0401
1.76	0.9608	0.0392
1.77	0.9616	0.0384
1.78	0.9625	0.0375
1.79	0.9633	0.0367
1.80	0.9641	0.0359
1.81	0.9649	0.0351
1.82	0.9656	0.0344
1.83	0.9664	0.0336
1.84	0.9671	0.0329
1.85	0.9678	0.0322
1.86	0.9686	0.0314
1.87	0.9693	0.0307

z Score	Body	Tail
1.88	0.9699	0.0301
1.89	0.9706	0.0294
1.90	0.9713	0.0287
1.91	0.9719	0.0281
1.92	0.9726	0.0274
1.93	0.9732	0.0268
1.94	0.9738	0.0262
1.95	0.9744	0.0256
1.96	0.9750	0.0250
1.97	0.9756	0.0244
1.98	0.9761	0.0239
1.99	0.9767	0.0233
2.00	0.9772	0.0228
2.01	0.9778	0.0222
2.02	0.9783	0.0217
2.03	0.9788	0.0212
2.04	0.9793	0.0207
2.05	0.9798	0.0202
2.06	0.9803	0.0197
2.07	0.9808	0.0192
2.08	0.9812	0.0188
2.09	0.9817	0.0183
2.10	0.9821	0.0179
2.11	0.9826	0.0174
2.12	0.9830	0.0170
2.13	0.9834	0.0166
2.14	0.9838	0.0162
2.15	0.9842	0.0158
2.16	0.9846	0.0154
2.17	0.9850	0.0150
2.18	0.9854	0.0146
2.19	0.9857	0.0143
2.20	0.9861	0.0139
2.21	0.9864	0.0136
2.22	0.9868	0.0132
2.23	0.9871	0.0129
2.24	0.9875	0.0125
2.25	0.9878	0.0122
2.26	0.9881	0.0119
2.27	0.9884	0.0116
2.28	0.9887	0.0113
2.29	0.9890	0.0110
2.30	0.9893	0.0107
2.31	0.9896	0.0104
2.32	0.9898	0.0102
2.33	0.9901	0.0099
2.34	0.9904	0.0096
2.35	0.9906	0.0094
2.36	0.9909	0.0091
2.37	0.9911	0.0089
2.38	0.9913	0.0087
2.39	0.9916	0.0084
2.40	0.9918	0.0082
2.41	0.9920	0.0080
2.42	0.9922	0.0078
2.43	0.9925	0.0075
2.44	0.9927	0.0073
2.45	0.9929	0.0071
2.46	0.9931	0.0069
2.47	0.9932	0.0068
2.48	0.9934	0.0066
2.49	0.9936	0.0064
2.50	0.9938	0.0062
2.51	0.9940	0.0060
2.52	0.9941	0.0059
2.53	0.9943	0.0057
2.54	0.9945	0.0055
2.55	0.9946	0.0054
2.56	0.9948	0.0052
2.57	0.9949	0.0051
2.58	0.9951	0.0049
2.59	0.9952	0.0048
2.60	0.9953	0.0047
2.61	0.9955	0.0045
2.62	0.9956	0.0044
2.63	0.9957	0.0043
2.64	0.9959	0.0041
2.65	0.9960	0.0040
2.66	0.9961	0.0039
2.67	0.9962	0.0038
2.68	0.9963	0.0037
2.69	0.9964	0.0036
2.70	0.9965	0.0035
2.71	0.9966	0.0034
2.72	0.9967	0.0033
2.73	0.9968	0.0032
2.74	0.9969	0.0031
2.75	0.9970	0.0030
2.76	0.9971	0.0029
2.77	0.9972	0.0028
2.78	0.9973	0.0027
2.79	0.9974	0.0026
2.80	0.9974	0.0026
2.81	0.9975	0.0025
2.82	0.9976	0.0024
2.83	0.9977	0.0023
2.84	0.9977	0.0023
2.85	0.9978	0.0022
2.86	0.9979	0.0021
2.87	0.9979	0.0021
2.88	0.9980	0.0020
2.89	0.9981	0.0019
2.90	0.9981	0.0019
2.91	0.9982	0.0018

(Continued)

(Continued)

z Score	Body	*Tail*	*z* Score	Body	*Tail*	*z* Score	Body	*Tail*	*z* Score	Body	*Tail*
2.92	0.9982	*0.0018*	**3.07**	0.9989	*0.0011*	**3.22**	0.9994	*0.0006*	**3.37**	0.9996	*0.0004*
2.93	0.9983	*0.0017*	**3.08**	0.9990	*0.0010*	**3.23**	0.9994	*0.0006*	**3.38**	0.9996	*0.0004*
2.94	0.9984	*0.0016*	**3.09**	0.9990	*0.0010*	**3.24**	0.9994	*0.0006*	**3.39**	0.9997	*0.0003*
2.95	0.9984	*0.0016*	**3.10**	0.9990	*0.0010*	**3.25**	0.9994	*0.0006*	**3.40**	0.9997	*0.0003*
2.96	0.9985	*0.0015*	**3.11**	0.9991	*0.0009*	**3.26**	0.9994	*0.0006*	**3.41**	0.9997	*0.0003*
2.97	0.9985	*0.0015*	**3.12**	0.9991	*0.0009*	**3.27**	0.9995	*0.0005*	**3.42**	0.9997	*0.0003*
2.98	0.9986	*0.0014*	**3.13**	0.9991	*0.0009*	**3.28**	0.9995	*0.0005*	**3.43**	0.9997	*0.0003*
2.99	0.9986	*0.0014*	**3.14**	0.9992	*0.0008*	**3.29**	0.9995	*0.0005*	**3.44**	0.9997	*0.0003*
3.00	0.9987	*0.0013*	**3.15**	0.9992	*0.0008*	**3.30**	0.9995	*0.0005*	**3.45**	0.9997	*0.0003*
3.01	0.9987	*0.0013*	**3.16**	0.9992	*0.0008*	**3.31**	0.9995	*0.0005*	**3.46**	0.9997	*0.0003*
3.02	0.9987	*0.0013*	**3.17**	0.9992	*0.0008*	**3.32**	0.9995	*0.0005*	**3.47**	0.9997	*0.0003*
3.03	0.9988	*0.0012*	**3.18**	0.9993	*0.0007*	**3.33**	0.9996	*0.0004*	**3.48**	0.9997	*0.0003*
3.04	0.9988	*0.0012*	**3.19**	0.9993	*0.0007*	**3.34**	0.9996	*0.0004*	**3.49**	0.9998	*0.0002*
3.05	0.9989	*0.0011*	**3.20**	0.9993	*0.0007*	**3.35**	0.9996	*0.0004*	**3.50**	0.9998	*0.0002*
3.06	0.9989	*0.0011*	**3.21**	0.9993	*0.0007*	**3.36**	0.9996	*0.0004*	**3.51**	0.9998	*0.0002*

Appendix B

One-Tailed Probabilities t Table

df	$\alpha = .05$	$\alpha = .01$
1	6.3138	31.8205
2	2.9200	6.9646
3	2.3534	4.5407
4	2.1318	3.7469
5	2.0150	3.3649
6	1.9432	3.1427
7	1.8946	2.9980
8	1.8595	2.8965
9	1.8331	2.8214
10	1.8125	2.7638
11	1.7959	2.7181
12	1.7823	2.6810
13	1.7709	2.6503
14	1.7613	2.6245
15	1.7531	2.6025
16	1.7459	2.5835
17	1.7396	2.5669
18	1.7341	2.5524
19	1.7291	2.5395
20	1.7247	2.5280
21	1.7207	2.5176
22	1.7171	2.5083
23	1.7139	2.4999
24	1.7109	2.4922
25	1.7081	2.4851
26	1.7056	2.4786
27	1.7033	2.4727
28	1.7011	2.4671
29	1.6991	2.4620
30	1.6973	2.4573
31	1.6955	2.4528
32	1.6939	2.4487
33	1.6924	2.4448
34	1.6909	2.4411
35	1.6896	2.4377
36	1.6883	2.4345
37	1.6871	2.4314
38	1.6860	2.4286
39	1.6849	2.4258
40	1.6839	2.4233
41	1.6829	2.4208
42	1.6820	2.4185
43	1.6811	2.4163
44	1.6802	2.4141
45	1.6794	2.4121
46	1.6787	2.4102
47	1.6779	2.4083
48	1.6772	2.4066
49	1.6766	2.4049
50	1.6759	2.4033
51	1.6753	2.4017
52	1.6747	2.4002
53	1.6741	2.3988
54	1.6736	2.3974
55	1.6730	2.3961
56	1.6725	2.3948
57	1.6720	2.3936
58	1.6716	2.3924
59	1.6711	2.3912
60	1.6706	2.3901
61	1.6702	2.3890
62	1.6698	2.3880
63	1.6694	2.3870
64	1.6690	2.3860
65	1.6686	2.3851
66	1.6683	2.3842
67	1.6679	2.3833
68	1.6676	2.3824
69	1.6672	2.3816
70	1.6669	2.3808
71	1.6666	2.3800
72	1.6663	2.3793
73	1.6660	2.3785
74	1.6657	2.3778
75	1.6654	2.3771
76	1.6652	2.3764
77	1.6649	2.3758
78	1.6646	2.3751
79	1.6644	2.3745
80	1.6641	2.3739
81	1.6639	2.3733
82	1.6636	2.3727
83	1.6634	2.3721
84	1.6632	2.3716
85	1.6630	2.3710
86	1.6628	2.3705
87	1.6626	2.3700
88	1.6624	2.3695
89	1.6622	2.3690
90	1.6620	2.3685
91	1.6618	2.3680
92	1.6616	2.3676
93	1.6614	2.3671
94	1.6612	2.3667
95	1.6611	2.3662
96	1.6609	2.3658
97	1.6607	2.3654
98	1.6606	2.3650
99	1.6604	2.3646
100	1.6602	2.3642
101	1.6601	2.3638
102	1.6599	2.3635
103	1.6598	2.3631
104	1.6596	2.3627
105	1.6595	2.3624
106	1.6594	2.3620
107	1.6592	2.3617
108	1.6591	2.3614
109	1.6590	2.3610
110	1.6588	2.3607
111	1.6587	2.3604
112	1.6586	2.3601
113	1.6585	2.3598
114	1.6583	2.3595
115	1.6582	2.3592
116	1.6581	2.3589
117	1.6580	2.3586
118	1.6579	2.3584
119	1.6578	2.3581
120	1.6577	2.3578
∞	1.6450	2.3260

Two-Tailed Probabilities t Table

df	α = .05	α = .01
1	12.7062	63.6567
2	4.3027	9.9248
3	3.1824	5.8409
4	2.7764	4.6041
5	2.5706	4.0321
6	2.4469	3.7074
7	2.3646	3.4995
8	2.3060	3.3554
9	2.2622	3.2498
10	2.2281	3.1693
11	2.2010	3.1058
12	2.1788	3.0545
13	2.1604	3.0123
14	2.1448	2.9768
15	2.1314	2.9467
16	2.1199	2.9208
17	2.1098	2.8982
18	2.1009	2.8784
19	2.0930	2.8609
20	2.0860	2.8453
21	2.0796	2.8314
22	2.0739	2.8188
23	2.0687	2.8073
24	2.0639	2.7969
25	2.0595	2.7874
26	2.0555	2.7787
27	2.0518	2.7707
28	2.0484	2.7633
29	2.0452	2.7564
30	2.0423	2.7500
31	2.0395	2.7440
32	2.0369	2.7385
33	2.0345	2.7333
34	2.0322	2.7284
35	2.0301	2.7238
36	2.0281	2.7195
37	2.0262	2.7154
38	2.0244	2.7116
39	2.0227	2.7079
40	2.0211	2.7045
41	2.0195	2.7012
42	2.0181	2.6981
43	2.0167	2.6951
44	2.0154	2.6923
45	2.0141	2.6896
46	2.0129	2.6870
47	2.0117	2.6846
48	2.0106	2.6822
49	2.0096	2.6800
50	2.0086	2.6778
51	2.0076	2.6757
52	2.0066	2.6737
53	2.0057	2.6718
54	2.0049	2.6700
55	2.0040	2.6682
56	2.0032	2.6665
57	2.0025	2.6649
58	2.0017	2.6633
59	2.0010	2.6618
60	2.0003	2.6603
61	1.9996	2.6589
62	1.9990	2.6575
63	1.9983	2.6561
64	1.9977	2.6549
65	1.9971	2.6536
66	1.9966	2.6524
67	1.9960	2.6512
68	1.9955	2.6501
69	1.9949	2.6490
70	1.9944	2.6479
71	1.9939	2.6469
72	1.9935	2.6459
73	1.9930	2.6449
74	1.9925	2.6439
75	1.9921	2.6430
76	1.9917	2.6421
77	1.9913	2.6412
78	1.9908	2.6403
79	1.9905	2.6395
80	1.9901	2.6387
81	1.9897	2.6379
82	1.9893	2.6371
83	1.9890	2.6364
84	1.9886	2.6356
85	1.9883	2.6349
86	1.9879	2.6342
87	1.9876	2.6335
88	1.9873	2.6329
89	1.9870	2.6322
90	1.9867	2.6316
91	1.9864	2.6309
92	1.9861	2.6303
93	1.9858	2.6297
94	1.9855	2.6291
95	1.9853	2.6286
96	1.9850	2.6280
97	1.9847	2.6275
98	1.9845	2.6269
99	1.9842	2.6264
100	1.9840	2.6259
101	1.9837	2.6254
102	1.9835	2.6249
103	1.9833	2.6244
104	1.9830	2.6239
105	1.9828	2.6235
106	1.9826	2.6230
107	1.9824	2.6226
108	1.9822	2.6221
109	1.9820	2.6217
110	1.9818	2.6213
111	1.9816	2.6208
112	1.9814	2.6204
113	1.9812	2.6200
114	1.9810	2.6196
115	1.9808	2.6193
116	1.9806	2.6189
117	1.9804	2.6185
118	1.9803	2.6181
119	1.9801	2.6178
120	1.9799	2.6174
∞	1.9600	2.5760

Appendix C

F *Table* ($\alpha = .05$)

	df Numerator									
***df* Denominator**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16

(Continued)

(Continued)

	df Numerator									
***df* Denominator**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
31	4.16	3.30	2.91	2.68	2.52	2.41	2.32	2.25	2.20	2.15
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14
33	4.14	3.28	2.89	2.66	2.50	2.39	2.30	2.23	2.18	2.13
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11
36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15	2.11
37	4.11	3.25	2.86	2.63	2.47	2.36	2.27	2.20	2.14	2.10
38	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09
39	4.09	3.24	2.85	2.61	2.46	2.34	2.26	2.19	2.13	2.08
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
55	4.02	3.16	2.77	2.54	2.38	2.27	2.18	2.11	2.06	2.01
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
65	3.99	3.14	2.75	2.51	2.36	2.24	2.15	2.08	2.03	1.98
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02	1.97
75	3.97	3.12	2.73	2.49	2.34	2.22	2.13	2.06	2.01	1.96
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00	1.95
85	3.95	3.10	2.71	2.48	2.32	2.21	2.12	2.05	1.99	1.94
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
95	3.94	3.09	2.70	2.47	2.31	2.20	2.11	2.04	1.98	1.93
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93
105	3.93	3.08	2.69	2.46	2.30	2.19	2.10	2.03	1.97	1.92
110	3.93	3.08	2.69	2.45	2.30	2.18	2.09	2.02	1.97	1.92
115	3.92	3.08	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91

F *Table* ($\alpha = .01$)

	df Numerator									
***df* Denominator**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
31	7.53	5.36	4.48	3.99	3.67	3.45	3.28	3.15	3.04	2.96
32	7.50	5.34	4.46	3.97	3.65	3.43	3.26	3.13	3.02	2.93
33	7.47	5.31	4.44	3.95	3.63	3.41	3.24	3.11	3.00	2.91
34	7.44	5.29	4.42	3.93	3.61	3.39	3.22	3.09	2.98	2.89
35	7.42	5.27	4.40	3.91	3.59	3.37	3.20	3.07	2.96	2.88
36	7.40	5.25	4.38	3.89	3.57	3.35	3.18	3.05	2.95	2.86
37	7.37	5.23	4.36	3.87	3.56	3.33	3.17	3.04	2.93	2.84
38	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.92	2.83
39	7.33	5.19	4.33	3.84	3.53	3.30	3.14	3.01	2.90	2.81
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
45	7.23	5.11	4.25	3.77	3.45	3.23	3.07	2.94	2.83	2.74

(Continued)

(Continued)

	df Numerator									
***df* Denominator**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70
55	7.12	5.01	4.16	3.68	3.37	3.15	2.98	2.85	2.75	2.66
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
65	7.04	4.95	4.10	3.62	3.31	3.09	2.93	2.80	2.69	2.61
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67	2.59
75	6.99	4.90	4.05	3.58	3.27	3.05	2.89	2.76	2.65	2.57
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64	2.55
85	6.94	4.86	4.02	3.55	3.24	3.02	2.86	2.73	2.62	2.54
90	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52
95	6.91	4.84	3.99	3.52	3.22	3.00	2.83	2.70	2.60	2.51
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50
105	6.88	4.81	3.97	3.50	3.20	2.98	2.81	2.69	2.58	2.49
110	6.87	4.80	3.96	3.49	3.19	2.97	2.81	2.68	2.57	2.49
115	6.86	4.79	3.96	3.49	3.18	2.96	2.80	2.67	2.57	2.48
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47

Appendix D

The Studentized Range Statistic (q) Table

alpha = .05 (*top*)
alpha = .01 (*bottom*)

	Number of Treatment Conditions					
***df* for Error Term**	**2**	**3**	**4**	**5**	**6**	**7**
5	3.64 5.70	4.60 6.98	5.22 7.80	5.67 8.42	6.03 8.91	6.33 9.32
6	3.46 5.24	4.34 6.33	4.90 7.03	5.30 7.56	5.63 7.97	5.90 8.32
7	3.34 4.95	4.16 5.92	4.68 6.54	5.06 7.01	5.36 7.37	5.61 7.68
8	3.26 4.75	4.04 5.64	4.53 6.20	4.89 6.62	5.17 6.96	5.40 7.24
9	3.20 4.60	3.95 5.43	4.41 5.96	4.76 6.35	5.02 6.66	5.24 6.91
10	3.15 4.48	3.88 5.27	4.33 5.77	4.65 6.14	4.91 6.43	5.12 6.67
11	3.11 4.39	3.82 5.15	4.26 5.62	4.57 5.97	4.82 6.25	5.03 6.48
12	3.08 4.32	3.77 5.05	4.20 5.50	4.51 5.84	4.75 6.10	4.95 6.32
13	3.06 4.26	3.73 4.96	4.15 5.40	4.45 5.73	4.69 5.98	4.88 6.19
14	3.03 4.21	3.70 4.89	4.11 5.32	4.41 5.63	4.64 5.88	4.83 6.08
15	3.01 4.17	3.67 4.84	4.08 5.25	4.37 5.56	4.59 5.80	4.78 5.99
16	3.00 4.13	3.65 4.79	4.05 5.19	4.33 5.49	4.56 5.72	4.74 5.92
17	2.98 4.10	3.63 4.74	4.02 5.14	4.30 5.43	4.52 5.66	4.70 5.85
18	2.97 4.07	3.61 4.70	4.00 5.09	4.28 5.38	4.49 5.60	4.67 5.79

(Continued)

(Continued)

	Number of Treatment Conditions					
***df* for Error Term**	**2**	**3**	**4**	**5**	**6**	**7**
19	2.96 4.05	3.59 4.67	3.98 5.05	4.25 5.33	4.47 5.55	4.65 5.73
20	2.95 4.02	3.58 4.64	3.96 5.02	4.23 5.29	4.45 5.51	4.62 5.69
30	2.89 3.89	3.49 4.45	3.85 4.80	4.10 5.05	4.30 5.24	4.46 5.40
40	2.86 3.82	3.44 4.37	3.79 4.70	4.04 4.93	4.23 5.11	4.39 5.26
60	2.83 3.76	3.40 4.28	3.74 4.59	3.98 4.82	4.16 4.99	4.31 5.13
120	2.80 3.70	3.36 4.20	3.68 4.50	3.92 4.71	4.10 4.87	4.24 5.01
Infinity	2.77 3.64	3.31 4.12	3.63 4.40	3.86 4.60	4.03 4.76	4.17 4.88

Source: Adapted from http://www.stat.duke.edu/courses/Spring98/sta110c/qtable.htm

Appendix E

One-Tailed Pearson Correlation Table

df	$\alpha = .05$	$\alpha = .01$	*df*	$\alpha = .05$	$\alpha = .01$	*df*	$\alpha = .05$	$\alpha = .01$	*df*	$\alpha = .05$	$\alpha = .01$
1	0.988	1.000	31	0.291	0.403	61	0.209	0.293	91	0.172	0.241
2	0.900	0.980	32	0.287	0.397	62	0.207	0.290	92	0.171	0.240
3	0.805	0.934	33	0.283	0.392	63	0.206	0.288	93	0.170	0.238
4	0.729	0.882	34	0.279	0.386	64	0.204	0.286	94	0.169	0.237
5	0.669	0.833	35	0.275	0.381	65	0.203	0.284	95	0.168	0.236
6	0.621	0.789	36	0.271	0.376	66	0.201	0.282	96	0.167	0.235
7	0.582	0.750	37	0.267	0.371	67	0.200	0.280	97	0.166	0.234
8	0.549	0.715	38	0.264	0.367	68	0.198	0.278	98	0.165	0.232
9	0.521	0.685	39	0.260	0.362	69	0.197	0.276	99	0.165	0.231
10	0.497	0.658	40	0.257	0.358	70	0.195	0.274	100	0.164	0.230
11	0.476	0.634	41	0.254	0.354	71	0.194	0.272	101	0.163	0.229
12	0.458	0.612	42	0.251	0.350	72	0.193	0.270	102	0.162	0.228
13	0.441	0.592	43	0.248	0.346	73	0.191	0.268	103	0.161	0.227
14	0.426	0.574	44	0.246	0.342	74	0.190	0.266	104	0.161	0.226
15	0.412	0.558	45	0.243	0.338	75	0.189	0.265	105	0.160	0.225
16	0.400	0.543	46	0.240	0.335	76	0.188	0.263	106	0.159	0.224
17	0.389	0.529	47	0.238	0.331	77	0.186	0.261	107	0.158	0.223
18	0.378	0.516	48	0.235	0.328	78	0.185	0.260	108	0.158	0.222
19	0.369	0.503	49	0.233	0.325	79	0.184	0.258	109	0.157	0.221
20	0.360	0.492	50	0.231	0.322	80	0.183	0.257	110	0.156	0.220
21	0.352	0.482	51	0.228	0.319	81	0.182	0.255	111	0.156	0.219
22	0.344	0.472	52	0.226	0.316	82	0.181	0.253	112	0.155	0.218
23	0.337	0.462	53	0.224	0.313	83	0.180	0.252	113	0.154	0.217
24	0.330	0.453	54	0.222	0.310	84	0.179	0.251	114	0.153	0.216
25	0.323	0.445	55	0.220	0.307	85	0.178	0.249	115	0.153	0.215
26	0.317	0.437	56	0.218	0.305	86	0.176	0.248	116	0.152	0.214
27	0.311	0.430	57	0.216	0.302	87	0.175	0.246	117	0.152	0.213
28	0.306	0.423	58	0.214	0.300	88	0.174	0.245	118	0.151	0.212
29	0.301	0.416	59	0.213	0.297	89	0.174	0.244	119	0.150	0.211
30	0.296	0.409	60	0.211	0.295	90	0.173	0.242	120	0.150	0.210

Two-Tailed Pearson Correlation Table

Two Tailed alpha

df	0.05	0.01	*df*	0.05	0.01	*df*	0.05	0.01
1	0.997	1.000	41	0.301	0.389	81	0.216	0.281
2	0.950	0.990	42	0.297	0.384	82	0.215	0.280
3	0.878	0.959	43	0.294	0.380	83	0.213	0.278
4	0.811	0.917	44	0.291	0.376	84	0.212	0.276
5	0.754	0.875	45	0.288	0.372	85	0.211	0.275
6	0.707	0.834	46	0.285	0.368	86	0.210	0.273
7	0.666	0.798	47	0.282	0.365	87	0.208	0.272
8	0.632	0.765	48	0.279	0.361	88	0.207	0.270
9	0.602	0.735	49	0.276	0.358	89	0.206	0.269
10	0.576	0.708	50	0.273	0.354	90	0.205	0.267
11	0.553	0.684	51	0.271	0.351	91	0.204	0.266
12	0.532	0.661	52	0.268	0.348	92	0.203	0.264
13	0.514	0.641	53	0.266	0.345	93	0.202	0.263
14	0.497	0.623	54	0.263	0.341	94	0.201	0.262
15	0.482	0.606	55	0.261	0.339	95	0.200	0.260
16	0.468	0.590	56	0.259	0.336	96	0.199	0.259
17	0.456	0.575	57	0.256	0.333	97	0.198	0.258
18	0.444	0.561	58	0.254	0.330	98	0.197	0.256
19	0.433	0.549	59	0.252	0.327	99	0.196	0.255
20	0.423	0.537	60	0.250	0.325	100	0.195	0.254
21	0.413	0.526	61	0.248	0.322	101	0.194	0.253
22	0.404	0.515	62	0.246	0.320	102	0.193	0.252
23	0.396	0.505	63	0.244	0.317	103	0.192	0.250
24	0.388	0.496	64	0.242	0.315	104	0.191	0.249
25	0.381	0.487	65	0.240	0.313	105	0.190	0.248
26	0.374	0.479	66	0.239	0.310	106	0.189	0.247
27	0.367	0.471	67	0.237	0.308	107	0.188	0.246
28	0.361	0.463	68	0.235	0.306	108	0.187	0.245
29	0.355	0.456	69	0.234	0.304	109	0.187	0.244
30	0.349	0.449	70	0.232	0.302	110	0.186	0.242
31	0.344	0.442	71	0.230	0.300	111	0.185	0.241
32	0.339	0.436	72	0.229	0.298	112	0.184	0.240
33	0.334	0.430	73	0.227	0.296	113	0.183	0.239
34	0.329	0.424	74	0.226	0.294	114	0.182	0.238
35	0.325	0.418	75	0.224	0.292	115	0.182	0.237
36	0.320	0.413	76	0.223	0.290	116	0.181	0.236
37	0.316	0.408	77	0.221	0.288	117	0.180	0.235
38	0.312	0.403	78	0.220	0.286	118	0.179	0.234
39	0.308	0.398	79	0.219	0.285	119	0.179	0.233
40	0.304	0.393	80	0.217	0.283	120	0.178	0.232

Appendix F

Spearman's Correlation Table

One-tail	0.05	0.025	0.01	0.005
Two-tail	0.1	0.05	0.02	0.01
$N = 4$	1.000			
5	0.900	1.000	1.000	
6	0.829	0.886	0.943	1.000
7	0.714	0.786	0.893	0.929
8	0.643	0.738	0.833	0.881
9	0.600	0.700	0.783	0.833
10	0.564	0.648	0.745	0.794
11	0.536	0.618	0.709	0.755
12	0.503	0.587	0.671	0.727
13	0.484	0.560	0.648	0.703
14	0.464	0.538	0.622	0.675
15	0.443	0.521	0.604	0.654
16	0.429	0.503	0.582	0.635
17	0.414	0.485	0.566	0.615
18	0.401	0.472	0.550	0.600
19	0.391	0.460	0.535	0.584
20	0.380	0.447	0.520	0.570
21	0.370	0.435	0.508	0.556
22	0.361	0.425	0.496	0.544
23	0.353	0.415	0.486	0.532
24	0.344	0.406	0.476	0.521
25	0.337	0.398	0.466	0.511
26	0.331	0.390	0.457	0.501
27	0.324	0.382	0.448	0.491
28	0.317	0.375	0.440	0.483
29	0.312	0.368	0.433	0.475

	0.05	0.025	0.01	0.005
	0.1	0.05	0.02	0.01
$N = 30$	0.306	0.362	0.425	0.467
31	0.301	0.356	0.418	0.459
32	0.296	0.350	0.412	0.452
33	0.291	0.345	0.405	0.446
34	0.278	0.340	0.399	0.439
35	0.283	0.335	0.394	0.433
36	0.279	0.330	0.388	0.427
37	0.275	0.325	0.383	0.421
38	0.271	0.321	0.378	0.415
39	0.267	0.317	0.373	0.410
40	0.264	0.313	0.368	0.405
41	0.261	0.309	0.364	0.400
42	0.257	0.305	0.359	0.395
43	0.254	0.301	0.355	0.391
44	0.251	0.298	0.351	0.386
45	0.248	0.294	0.347	0.382
46	0.246	0.291	0.343	0.378
47	0.243	0.288	0.340	0.374
48	0.240	0.285	0.336	0.370
49	0.238	0.282	0.333	0.366
50	0.235	0.279	0.329	0.363
60	0.214	0.255	0.300	0.331
70	0.190	0.235	0.278	0.307
80	0.185	0.220	0.260	0.287
90	0.174	0.207	0.245	0.271
100	0.165	0.197	0.233	0.257

Source: Adapted from http://www.oup.com/uk/orc/bin/9780199265152/01student/04calculation/chapter_11_spearman.xls

Appendix G

Fisher r to z Table

r	z
0.01	0.0100
0.02	0.0200
0.03	0.0300
0.04	0.0400
0.05	0.0500
0.06	0.0601
0.07	0.0701
0.08	0.0802
0.09	0.0902
0.1	0.1003
0.11	0.1104
0.12	0.1206
0.13	0.1307
0.14	0.1409
0.15	0.1511
0.16	0.1614
0.17	0.1717
0.18	0.1820
0.19	0.1923
0.2	0.2027
0.21	0.2132
0.22	0.2237
0.23	0.2342
0.24	0.2448
0.25	0.2554
0.26	0.2661
0.27	0.2769
0.28	0.2877
0.29	0.2986
0.3	0.3095
0.31	0.3205
0.32	0.3316
0.33	0.3428
0.34	0.3541
0.35	0.3654
0.36	0.3769
0.37	0.3884
0.38	0.4001
0.39	0.4118
0.4	0.4236
0.41	0.4356
0.42	0.4477
0.43	0.4599
0.44	0.4722
0.45	0.4847
0.46	0.4973
0.47	0.5101
0.48	0.5230
0.49	0.5361
0.5	0.5493
0.51	0.5627
0.52	0.5763
0.53	0.5901
0.54	0.6042
0.55	0.6184
0.56	0.6328
0.57	0.6475
0.58	0.6625
0.59	0.6777
0.6	0.6931
0.61	0.7089
0.62	0.7250
0.63	0.7414
0.64	0.7582
0.65	0.7753
0.66	0.7928
0.67	0.8107
0.68	0.8291
0.69	0.8480
0.7	0.8673
0.71	0.8872
0.72	0.9076
0.73	0.9287
0.74	0.9505
0.75	0.9730
0.76	0.9962
0.77	1.0203
0.78	1.0454
0.79	1.0714
0.8	1.0986
0.81	1.1270
0.82	1.1568
0.83	1.1881
0.84	1.2212
0.85	1.2562
0.86	1.2933
0.87	1.3331
0.88	1.3758
0.89	1.4219
0.9	1.4722
0.91	1.5275
0.92	1.5890
0.93	1.6584
0.94	1.7380
0.95	1.8318
0.96	1.9459
0.97	2.0923
0.98	2.2976
0.99	2.6467

Appendix H

Critical Values for Chi-Square

df	$\alpha = 0.05$	$\alpha = 0.01$
1	3.841	6.635
2	5.991	9.210
3	7.815	11.345
4	9.488	13.277
5	11.070	15.086
6	12.592	16.812
7	14.067	18.475
8	15.507	20.090
9	16.919	21.666
10	18.307	23.209
11	19.675	24.725
12	21.026	26.217
13	22.362	27.688
14	23.685	29.141
15	24.996	30.578
16	26.296	32.000
17	27.587	33.409
18	28.869	34.805
19	30.144	36.191
20	31.410	37.566
21	32.671	38.932
22	33.924	40.289
23	35.172	41.638
24	36.415	42.980
25	37.652	44.314
26	38.885	45.642
27	40.113	46.963
28	41.337	48.278
29	42.557	49.588
30	43.773	50.892

Appendix I

Computing SSs for Factorial ANOVA

Low Body Temperature		High Body Temperature	
Beer	Placebo	Beer	Placebo
10	5	14	11
8	8	17	8
7	6	12	9
8	7	19	13
9	4	18	8
7	10	15	12
12	7	12	10
11	9	21	9
$\Sigma X = 72$; $\Sigma X^2 = 672$	$\Sigma X = 56$; $\Sigma X^2 = 420$	$\Sigma X = 128$; $\Sigma X^2 = 2{,}124$	$\Sigma X = 80$; $\Sigma X^2 = 824$

The above table displays the scores within each of the four conditions in this study. The ΣX and the ΣX^2 within each condition are displayed in each column of the table.

We will start by computing the *SS* for all of the scores by summing all of the scores to find $\Sigma X_{total} = 72 + 56 + 128 + 80 = 336$. Then, the sum of all the squared scores is $\Sigma X^2 = 672 + 420 + 2124 + 824 = 4{,}040$. Use the *SS* formula to find the SS_{total}, where N = the total number of scores or 32.

$$SS_{total} = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = 4040 - \frac{(336)^2}{32} = 512.$$

Then, compute the SS_{within} (also called the SS_{error}) by computing the individual *SS*s within each of the four conditions. The *SS* for the low body temperature, beer condition is computed as follows:

$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = 672 - \frac{(72)^2}{8} = 24.$$

By using the same procedure for the other three conditions, we find the other condition *SS*s to be 28, 76, and 24. The SS_{within} is the sum of the individual condition *SS*s, or 152.

$$SS_{within} = 24 + 28 + 76 + 24 = 152.$$

The next step is to compute the *SS*s for each of the three effects we are interested in—namely, the main effect of body temperature, the main effect of alcohol, and the interaction. First, we compute the *SS* for all of these effects combined, called the $SS_{between}$. This is done by simply subtracting the SS_{within} from the SS_{total}.

$$SS_{between} = 512 - 152 = 360.$$

Next, we must determine how much of the $SS_{between}$ variability was due to the main effect of Factor A—in this case, how much of the 360 variability was due to differences in body temperature. This is done by using the following formula:

$$\begin{aligned} SS_{\text{body temp}} &= \sum \frac{A_{\text{level}}^2}{n_{\text{level}}} - \frac{T^2}{N} \\ &= \frac{(72+56)^2}{16} + \frac{(128+80)^2}{16} - \frac{(72+56+128+80)^2}{32} \\ &= 1{,}024 + 2{,}704 - 3{,}528 = 200. \end{aligned}$$

A_{level} represents the sum of the scores in a given condition of Factor A. In the present case, Factor A is body temperature, which has two levels so that there will be two A_{level} terms in the equation. For example, the sums of the scores in the low body temperature conditions were 72 and 56, from the low body temperature with beer and low body temperature with placebo conditions, respectively. The sums of the scores in the high body temperature conditions were 128 and 80, respectively. In the above equation, *T* represents the sum of all the scores in the design, n_{level} represents the number of scores in a given level of Factor A, and *N* represents the number of scores in the entire study. So, the $SS_{body\ temp}$ accounts for 200 of the 360 between-treatment variability. A similar procedure yields the *SS* for alcohol.

$$\begin{aligned} SS_{\text{alcohol}} &= \sum \frac{A_{\text{level}}^2}{n_{\text{level}}} - \frac{T^2}{N} \\ &= \frac{(72+128)^2}{16} + \frac{(56+80)^2}{16} - \frac{(72+56+128+80)^2}{32} \\ &= 2{,}500 + 1{,}156 - 3{,}528 = 128. \end{aligned}$$

So the $SS_{alcohol}$ accounts for another 128 of the 360 between-treatment variability.

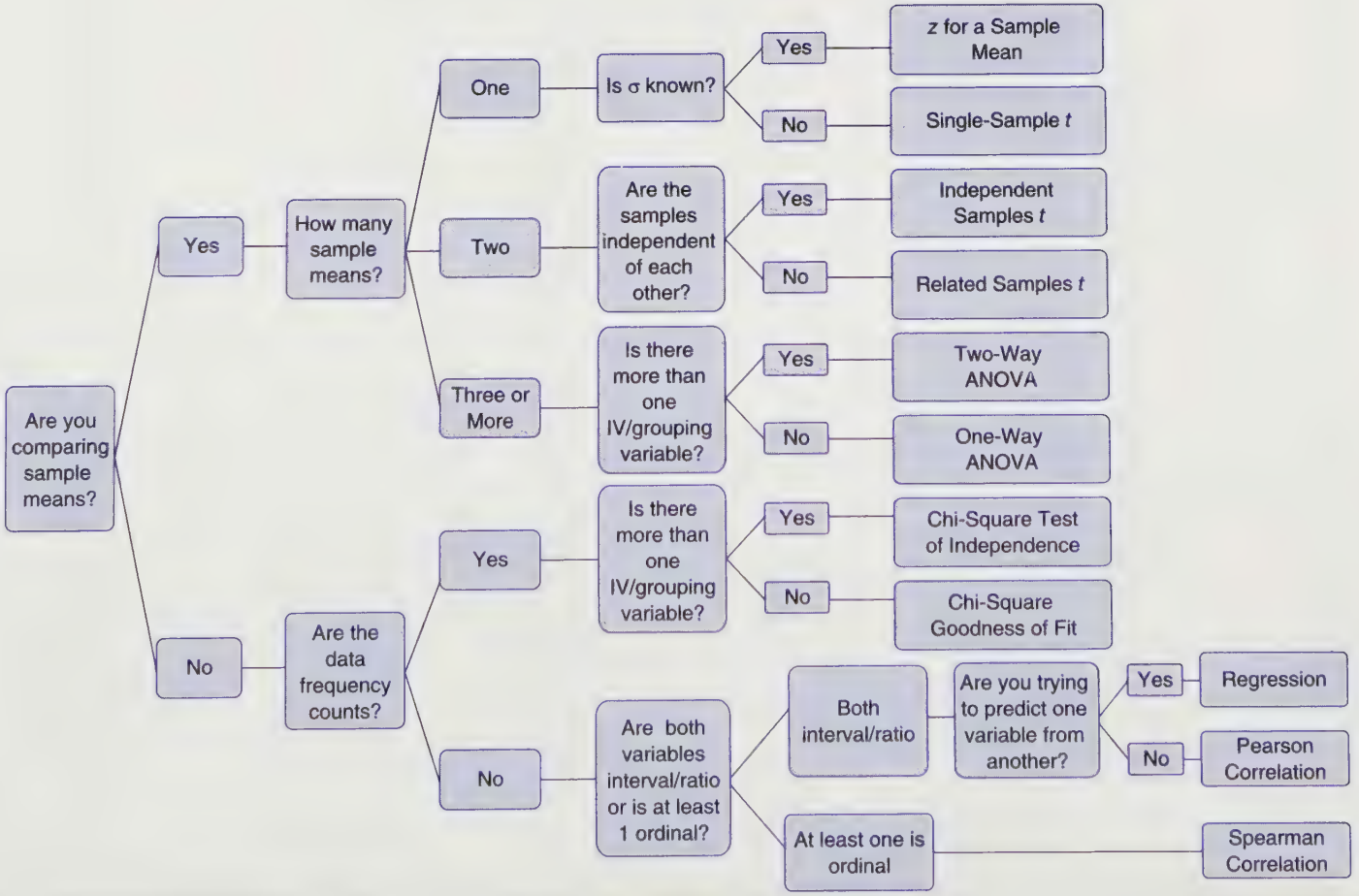
The final *SS* we need is the $SS_{interaction}$. As stated above, the three effects all created an $SS_{between}$ of 360. If the $SS_{body\ temp}$ was 200 and the $SS_{alcohol}$ was 128, the $SS_{interaction}$ must be 32.

$$SS_{interaction} = SS_{between} - SS_{body\ temp} - SS_{alcohol} = 360 - 200 - 128 = 32.$$

Appendix J

Choosing Correct Test Statistics

Decision Tree



Charts

Statistics for Comparing Means	
Goal	Statistic
Compare a sample mean to a population mean or other value	Use a z for a sample mean if σ is known Use a single-sample t if σ is unknown
Compare two sample means	Use an independent measures t if the two samples are independent of each other Use a related samples t if the two samples are related (either matched or repeated measures)
Compare three or more sample means	Use a one-way (single-factor) ANOVA if there is one independent variable Use a two-way (two-factor) ANOVA if there are two independent variables

Statistics for Doing Things Other Than Comparing Means	
Goal	Statistic
Analyze frequency counts	Use a chi-square goodness-of-fit test if there is one IV/grouping variable Use a chi-square test of independence if there are two IVs/grouping variables
Determine if two ordinal, interval, or ratio variables are related to each other	Use a Pearson correlation if both variables are interval/ratio and linearly related Use a Spearman correlation if at least one variable is ordinal or if the relationship is not linear (but is monotonic)
Use one interval/ratio variable to predict another interval/ratio variable	Use regression

Index

Abel, E. L., 354, 366
Algebra, in statistics, 3
Alpha (α) level:
  correlations, 535
  hypothesis testing rules, 290–291
  independent samples *t*, 330
  one-way independent samples ANOVA, 382–383
  single sample *t* test, 222–223
  statistical power and, 191–193
American College Test (ACT), 95–96
American Psychological Association (APA):
  confidence intervals reporting style, 249, 265–266
  correlation reporting style, 529
  significance tests, 6
  two-way ANOVA reporting style, 463–464
Analysis of variance. *See* ANOVA *entries*
ANOVA, one-way, 367–438
  between-treatment variability, 392–394
  confidence intervals, 423–427
  example problem, 371–372
  family-wise error and alpha inflation, 382–383
  F ratio, 394–395
  overview, 367–371
  post hoc tests, 395–398
  SPSS, 383–387
  test statistic selection, 428–430
  within-treatment variability (error), 389–392
ANOVA, one-way, activities:
  income inequality and politics, 490–493
  overview, 388–389
  SPSS for independent ANOVA, 400–410
  SPSS for one-way ANOVA, 398–400
  within- and between-group variability, 411–423
ANOVA, two-way, 439–512
  critical region, 450–453
  effect sizes, 454–457
  factorial designs, 440–441, 614–615
  logic of, 441–445
  null and research hypotheses, 446–450
  purpose of, 439–440
  results summarization, 457
  SPSS, 458–464
  statistical assumptions, 445–446
  test statistic, 452–454
ANOVA, two-way, activities:
  cell phone use while driving, 476–481
  income inequality and politics, 490–502
  non-drug treatments for depression, 481–489
  pharmaceuticals, 464–476
  test statistic choice, 502–505
APA format. *See* American Psychological Association (APA)
Appropriate measurement of variables, 146–147, 211, 219, 274
Armed Services Vocational Aptitude Battery (ASVAB), 233

Bar graphs:
  example of, 25–26
  overview, 13–14
  symmetrical and skewed distributions, 42
Baseball Register, 354
Bell-shaped distributions, 15–16
Beta (β) error, 185. *See also* Type I and type II error
Body temperature example of probabilities, 107–110

Causal hypothesis testing, 6–7
Causation, correlation *vs.*, 523–524
Cell means, in two-way ANOVA, 442
Cell phone use while driving, two-way ANOVA on, 476–481
Central limit theorem (CLT):
  correct statistic selection and, 140–141
  distribution of sample means and, 171
  overview, 136
  probability of given sample from, 139–140
  sampling error in, 137–139
  shape, center, and spread of distribution of sample means, 119–121
Central tendency, 39–63
  mean, 42–45
  median, 45–46
  mode, 47
  overview, 39–42
  SPSS on, 47–51
Central tendency, activities:
  deviation score computation, 50–60
  frequency distribution, 59–60
  graphing, 52–54
  mean, graphs of, 55–58
  mean, median, and mode computation, 51–52
  summary of, 60
Chi-square statistics, 567–596
  critical values table, 613
  logic of, 569–570
  overview, 567–569

Chi-square statistics, activities:
goodness-of-fit *vs.* chi-square for independence, 583–588
test statistic choice, 588–592
Chi-square statistics, for goodness-of-fit:
chi-square for independence *vs.*, 570
degrees of freedom and critical region, 571–572
null and research hypotheses, 571
results summarization, 573
SPSS, 578–579
statistical assumptions, 571
test statistic, 572–573
Chi-square statistics, for independence:
chi-square for goodness-of-fit *vs.*, 570
degrees of freedom and critical region, 574
effect size, 576–577
null and research hypotheses, 574
results summarization, 577
SPSS, 580–583
statistical assumptions, 573
test statistic, 574–576
CLT (central limit theorem). *See* Central limit theorem (CLT)
Coefficient of determination, 528
Coefficients of correlation, 513–514, 516–519. *See also* Correlation
Cohen, Jacob, 156, 206, 372n, 438, 573, 596
Cohen's *d*, 156, 169
Computational formulas:
correlation and regression, 519–520
SS, 68–69, 87
Condition means, in two-way ANOVA, 442
Confidence intervals, 241–270
APA style of reporting, 249, 265–266
description of, 5–6
correlations, 532–533, 546–549
effect sizes, 249–250
mean difference, 247–249
mean differences between independent samples, 357–361
one-way independent samples ANOVA, 423–427
population mean, 245–247
repeated/related samples *t* test, 299–308
interpretations of, 250–251
logic of, 244–245
purpose in statistical procedures, 241–243
sample means and sample mean difference, 252–266
SPSS, 251–252
Confounding variables, 241
Consumer Financial Protection Bureau, 211
Continuous variables, 12
Corley, K. M., 112,
Correlation:
causation *vs.*, 523–524
computational formulas for, 519–520
confidence intervals for, 532–533
correlation coefficient direction and strength, 516–519
hypothesis testing with, 524–525
logic of, 515–516
overview, 513–514
scatterplots for, 520–523
SPSS, 533–536
z scores and, 514–515
See also Pearson's correlation (*r*); Spearman's correlations (r_s)
Correlation, activities:
confidence intervals, 546–549
learning strategies example, 540–546
regression and prediction, 552–559
scatterplots, 536–539
Spearman's correlations (r_s), 550–551
test statistic choice, 560–562
Cramer's phi (ϕ), 576, 577
Critical region:
chi-square statistics for goodness-of-fit, 571–572
chi-square statistics for independence, 574
hypothesis testing with repeated/related samples *t* test, 290–291
hypothesis testing with *z* scores, 168, 171–172
one-tailed independent samples *t*, 327–328
one-tailed Pearson's correlation, 531
one-tailed repeated/related samples *t* test, 281, 291–292
one-tailed single sample *t* test, 213–214
one-tailed *vs.* two-tailed tests, 217–218
sample mean example (one-tailed) of hypothesis testing, 150–154
two-tailed independent samples *t*, 322
two-tailed Pearson's correlation, 527–528
two-tailed repeated/related samples *t* test, 276–277, 294
two-tailed single sample *t* test, 220
two-way ANOVA, 450–452
Critical value:
chi-square, 613
define critical region, 152–153
desired confidence interval, 245–246
F for one-way independent samples ANOVA, 374–375
p values, null hypothesis and, 177–181
Cumming, G., 249–250, 269, 279n, 314, 325n, 341n, 366
Cut lines, using *x* to find, 97–98

Data, graphing, 12–15
Definitional formulas, 68–69, 86, 519
Degrees of freedom (df):
chi-square statistics for goodness-of-fit, 571–572
chi-square statistics for independence, 574
one-tailed independent samples *t*, 327–328
one-tailed repeated/related samples *t* test, 281
one-tailed single sample *t* test, 213–214
sampling error associated with, 245
two-tailed independent samples *t*, 322
two-tailed repeated/related samples *t* test, 276–277
two-tailed single sample *t* test, 220

Dependent samples *t* test, 272
Dependent variables:
causal hypothesis testing, 6
description of, 8–9
two-way ANOVA, 439–440
Depression treatment, two-way ANOVA on (activity), 481–489
Descriptive statistics, 8, 49, 386
Deviation score, 59
DeWall, C. C., 336, 355, 366
Discrete variables, 12
Distributed practice, as study method, 540
Distributions:
building a distribution, 173–174
normality of, 146–147, 211, 219
shapes of, 15–19
See also Sample means, distribution of

Effect size:
chi-square statistics for independence, 576–577
confidence intervals for, 249–250
description of, 5–6, 169, 176–177
hypothesis testing with independent samples *t*, 340
one-tailed independent samples *t*, 329
one-tailed Pearson's correlation, 531
one-tailed repeated/related samples *t* test, 283, 292
one-tailed single sample *t* test, 216
one-way independent samples ANOVA, 380–382
overview, 195
purpose of, 243
repeated/related samples *t* test activity, 299–308
sample mean example (one-tailed) of hypothesis testing, 155–157
two-tailed independent samples t, 325
two-tailed Pearson's correlation, 528–529
two-tailed repeated/related samples *t* test, 279, 294–295
two-tailed single sample *t* test, 222
two-way ANOVA, 454–457
type I and type II error reduction impact on, 196–203
Estimation. *See* Confidence intervals
Expected frequency, in chi-square statistics, 569, 572, 574–575, 584
Experimental designs, 6, 284

Factorial designs, 440–441
Family-wise error, 382–383
Fei, Y. Y., 112
F ratio, 394–395
Frequency distribution table:
central tendency, 59
median identified in, 46
mode identified in, 47
overview, 19
SPSS, 21–24
F statistic. *See* ANOVA, one-way; ANOVA, two-way
F tables, 603–606

Gaussian distributions, 15–16
General Social Survey activity, 27–30
Goodness-of-fit. *See* Chi-square statistics
Graphing data:
central tendency measures, 52–55
mean, 55–58
overview, 12–15

Harris, G. T., 5, 37
Haub, Mark, 503
Helping professions, statistics in, 4–5
Heteroscedastic data, 522
Histograms:
example of, 23
overview, 13–15
z scores for normally shaped distribution, 100
Homogeneity of variance (Levene's test), 332–335
Homoscedastic data, 522
Honestly significant difference (HSD) test. *See* Tukey's honestly significant difference (HSD) test
Hypothesis testing:
causal, 6–7
correlation and regression, 524–525
description of, 5
effect size and confidence intervals, 5–6
statistical power and type I and type II error, 185–189
See also Null hypotheses; Research hypotheses
Hypothesis testing with independent samples *t*:
effect size, 340
example of, 336–337
null and research hypotheses, 337–340
overview, 336
results summarization, 340
significance testing summary, 344–345
statistical assumptions, 337
type I and II errors and statistical power, 341–344
Hypothesis testing with repeated/related samples *t* test:
one-tailed, 289–293
overview, 289
SPSS, 295–298
two-tailed, 293–295
Hypothesis testing with *z* scores, 145–206
critical values, *p* values, and null hypothesis activity, 177–181
effect size activity, 195–203
errors in, 158–160
"failure to reject the null," 164–165
overview, 145–146
p value, 162–163
"research suggests" meaning, 165
rules of, 160–162
sample mean example, 146–157
statistical power and type I and II error activity, 182–195
statistical significance in, 157–158
Hypothesis testing with *z* scores, activity:
critical region, 168, 171–172
effect size description, 169, 176–177

null and research hypotheses, 167–168, 170–172
results summarization, 169–170
statistical assumptions, 166–167
test statistic, 169
type I errors, type II errors, and statistical power, 173–176

Income inequality and politics, two-way ANOVA on, 490–502
Independence chi-square statistic. *See* Chi-square statistics
Independence of data, 146–147, 211, 219, 274
Independent research design (activity), 355–356
Independent samples *t*, 315–366
alpha levels for, 330
confidence intervals for mean differences between independent samples (activity), 357–361
formula for, 319
independent, matched, and related research designs (activity), 355–356
single-sample *t* and related samples *t vs.*, 315–319
SPSS, 330–335
test statistic choice (activity), 353–355
Independent samples *t*, hypothesis testing with (activity):
effect size, 340
example of, 336–337
null and research hypotheses, 337–340
overview, 336
results summarization, 340
significance testing summary, 344–345
statistical assumptions, 337
type I and II errors and statistical power, 341–344
Independent samples *t*, one-tailed:
degrees of freedom and critical region, 327–328
effect size, 329
null and research hypotheses, 326–327
results interpretation, 329–330
statistical assumptions, 326
test statistic, 328–329
Independent samples *t*, two-tailed:
activity, 347–353
degrees of freedom and critical region, 322
effect size, 325
null and research hypotheses, 321–322
results interpretation, 326
statistical assumptions, 320–321
test statistic, 322–325
Independent variables:
causal hypothesis testing, 6
description of, 8–9
two-way ANOVA, 439–440
Individual differences, variance affected by, 368, 370
Inferential statistics, 7, 113–114
Interaction effect, in two-way ANOVA, 443–444
Interaction effect size, 454–456
Interaction null and research hypotheses, in two-way ANOVA, 446–448
International Monetary Fund (IMF), 560
Interval scale of measurement, 10–11, 40–41

Johnson, M. E., 112

Kansas State University, 503
Kopp, M., 476, 512
Kruger, M. L., 354, 366
Kurtosis, in distributions, 18–19

Langer, P., 476, 512
Law of large numbers, 122
Leptokurtic distributions, 18–19
Levene's test, 321, 332–335
Levine, M. M., 107, 112
Linear scatterplots, 521–522, 537
Line graphs, 13, 15
Lower boundary of confidence intervals, 244, 247–248

Mackin, R., 227, 239
Mackowiak, P. A., 107, 112
Magnet, W., 476, 512
Main effect:
effect size for, 456–457
Factor A significance test, 442–443
Factor B significance test, 443
null and research hypotheses, 448–450
Marginal means for studying method, 449
Margin of error, 244, 547
Matched research design, 355–356
Matched samples *t* test, 272
Math skills required for statistics, 3–4
Mean:
central tendency measure, 42–45
computation of, 51–52
confidence interval for, 251–252
distribution of sample means, 117–118
graphing, 55–58
nominal scale of measurement, 41
SPSS to compare, 263–265
symmetrical distributions, 41–42
See also Sample mean example (one-tailed) of hypothesis testing; Sample means, distribution of
Mean difference, 247–249, 252–253
Measurement, 9–12, 30–31
Measurement error, 193, 368, 370
Median:
central tendency measure, 45–46
computation of, 51–52
ordinal scale of measurement, 39, 41
skewed distributions, 41–42
Mode, 47, 51–52
Monotonic scatterplots, 521–522, 537

Negatively skewed distributions, 16–19
Negative *z* scores, 102–103
NHST (null hypothesis significance testing), 5–6
Nominal scale of measurement, 10–11, 41
Nonmonotonic scatterplots, 522, 537

Normal distributions:
overview, 15–16
z scores for probability statements on, 98–103
Normality of distributions, 146–147, 211, 219, 274
Notation, mathematical, 4
Null hypotheses:
chi-square statistics for goodness-of-fit, 571
chi-square statistics for independence, 574
"failure to reject," 164–165
hypothesis testing with independent samples *t*, 337–340
hypothesis testing with *z* scores, 167–168, 170–172
one-tailed independent samples *t*, 326–327
one-tailed Pearson's correlation, 529–531
one-tailed repeated/related samples *t* test, 280–281, 291
one-tailed single sample *t* test, 212–213
one-way independent samples ANOVA, 373–374
p values, critical values and, 177–181
sample mean example (one-tailed) of hypothesis testing, 148–150
two-tailed independent samples *t*, 321–322
two-tailed Pearson's correlation, 527
two-tailed repeated/related samples *t* test, 275–276, 293
two-tailed single sample *t* test, 219–220
two-way ANOVA, 446–450
Null hypothesis significance testing (NHST), 5–6

Observed frequency, in chi-square statistics, 569, 572, 575
One-tailed independent samples *t*:
degrees of freedom and critical region, 327–328
effect size, 329
null and research hypotheses, 326–327
results interpretation, 329–330
statistical assumptions, 326
One-tailed Pearson *(r)* correlation:
critical region, 531
effect size, 531
null and research hypotheses, 529–530
overview, 529
results summarization, 531–532
statistical assumptions, 529
table of, 609
test statistic, 531
One-tailed probabilities *t* table, 601
One-tailed repeated/related samples *t* test:
degrees of freedom and critical region, 281
effect size, 283
hypothesis testing with, 289–293
null and research hypotheses, 280–281
results interpretation, 283
statistical assumptions, 280
test statistic, 281–283
One-tailed single sample *t* test:
activity on, 230–232
degrees of freedom and critical regions, 213–214
effect size, 216
null and research hypotheses, 212–213
results interpretation, 216–217
statistical assumptions, 211–212
test statistic, 214–216, 232–233
See also Single sample *t* test
One-way independent samples ANOVA. *See* ANOVA, one-way
Order of mathematical operations, 3–4
Ordinal scale of measurement, 10–11, 39, 41
Outliers, 40–41, 523

Paired samples *t* test, 272
PANAS (Positive and Negative Affect Scale), 297
Partial eta squared (η_p^2), 454
Pearson *(r)* correlation:
computing, 519–520
one-tailed, 529–532, 609
Spearman's correlation (r_s) *vs.*, 521–522, 537–539, 560
two-tailed, 610
Pharmaceuticals study, two-way ANOVA for, 464–476
Platykurtic distributions, 18, 19
Point estimates:
confidence intervals, 547
population mean, 244
Politics of income inequality, two-way ANOVA on, 490–502
Pooled variance, 328
Population:
confidence intervals for mean of, 245–247
distribution of sample means, 128–136
estimating parameter for, 261–263
mean of, 118, 154
point estimates of mean of, 244
sample as representative of, 5, 7–8
Population, variability in:
deviation score computation, 66–67
overview, 65–66
squaring deviation scores, 67–68
standard deviation computation, 69–72
summing squared deviation scores, 68–69
variance computation, 69
Positive and Negative Affect Scale (PANAS), 297
Positively skewed distributions, 16–17
Positive *z* scores, 100–102
Post hoc tests, in one-way independent samples ANOVA, 395–398
Practice retrieval, as study method, 540
Prediction, from regression, 552–559
"Pre-post" test, 272
Probabilities:
central limit theorem (CLT) and, 139–140
exact *vs.* estimates, 125–126
one-tailed *t*, table of, 601
replication recapture rate, 250
two-tailed *t*, table of, 602
z scores and (activity), 104–111
Proportion between two *z* scores, 103–104

p value:
critical values, null hypothesis and, 177–181
hypothesis testing with *z* scores, 162–163

Qualitative data from nominal measurement scale, 10
Quantitative data from ratio measurement scale, 10
Quinsey, V. L., 5

Rank ordering from ordinal measurement scale, 10
Ratio scale of measurement, 10–11, 40–41
Rausch, J. T., 107, 112
Regression, 552–559
foot length–height regression line, 554
regression equation, 553
SPSS, 554–555, 557
Related research design (activity), 355–356
Renner, M. J., 227, 239
Repeated/related samples *t* test, 271–314
hypothesis testing with (activity), 289–298
independent samples *t vs.*, 315–319
overview, 271–272
significance testing, effect sizes, and confidence intervals (activity), 299–308
single-sample *t* test and, 273–274
SPSS, 284–288
statistical results, experimental designs, and scientific conclusions, 284
Repeated/related samples *t* test, one-tailed:
degrees of freedom and critical region, 281
effect size, 283
null and research hypotheses, 280–281
results interpretation, 283
statistical assumptions, 280
test statistic, 281–283
Repeated/related samples *t* test, two-tailed:
degrees of freedom and critical region, 276–277
effect size, 279
null and research hypotheses, 275–276
results interpretation, 279–280
statistical assumptions, 274–275
test statistic, 277–279
Replication recapture rate, 250
Rereading chapters, as study method, 540
Research designs (activity), 355–356
Research hypotheses:
chi-square statistics for goodness-of-fit, 571
chi-square statistics for independence, 574
hypothesis testing with independent samples *t*, 337–340
hypothesis testing with *z* scores, 167–168, 170–172
one-tailed independent samples *t*, 326–327
one-tailed Pearson's correlation, 529–530
one-tailed repeated/related samples *t* test, 280–281, 291
one-tailed single sample *t* test, 212–213
one-way independent samples ANOVA, 373–374
sample mean example (one-tailed) of hypothesis testing, 148–150
two-tailed independent samples *t*, 321–322
two-tailed Pearson's correlation, 527
two-tailed repeated/related samples *t* test, 275–276, 293–294
two-tailed single sample *t* test, 219–220
two-way ANOVA, 446–450
"Research suggests," meaning of, 165
Results summarization:
APA style for confidence intervals, 249, 265–266
chi-square statistics for goodness-of-fit, 573
chi-square statistics for independence, 577
hypothesis testing with independent samples *t*, 340
hypothesis testing with *z* scores, 169–170
one-tailed independent samples *t*, 329–330
one-tailed Pearson's correlation, 531–532
one-tailed repeated/related samples *t* test, 283, 293
one-tailed single sample *t* test, 216–217
one-way independent samples ANOVA, 382
sample mean example (one-tailed) of hypothesis testing, 157
statistical power and type I and type II error activity, 194–195
two-tailed independent samples *t*, 326
two-tailed Pearson's correlation, 529
two-tailed repeated/related samples *t* test, 279–280, 295
two-tailed single sample *t* test, 222
two-way ANOVA, 457
Rice, M. E., 5, 37
Rosa, E., 25, 37
Rosa, L., 25, 37

Sample mean example (one-tailed) of hypothesis testing:
critical region, 150–154
effect size, 155–157
null and research hypothesis, 148–150
overview, 146
results interpretation, 157
statistical assumptions, 146–148
test statistic, 154–155
Sample means, distribution of, 113–144
activity on, 126–127
center of confidence interval, 244–245
central limit theorem and. *See* Central limit theorem (CLT)
description of, 115–117
estimation activity, 254–266
exact probabilities *vs.* probability estimates, 125–126
larger populations and samples (activity), 132–136
mean of, 117–118
sampling error and, 113–115
shape of, 119
smaller populations and samples (activity), 128–132
standard deviation of, 118–119
standard error of the mean in, 121–122
z for, 122–125

Samples:
distribution of sample means for larger (activity), 132–136
distribution of sample means for smaller (activity), 128–132
populations and, 5, 7–8
variability in, 72–74
Sample size:
sampling error reduced by, 115
t distribution shape and, 209–210
type I and type II error, statistical power and, 190–191
Sample statistic, description of, 7. *See also* Test statistic
Sampling error:
average mean difference expected, 278, 282
central limit theorem (CLT), 137–139
difference between sample and population, 5, 8
distribution of sample means and, 113–115
expected, 214–215, 245–246, 323–324
inferential statistics, 113–114
sample standard deviation to estimate, 207, 221
See also Standard error of the mean
Sarner, L., 25, 37
SAT (Scholastic Aptitude Test), 345
Scatterplots:
correlation appropriateness from, 537–539
correlation choice from, 526–527, 529, 531
correlation and regression, 520–523
correlation strength from, 536
direction and strength of relationships in, 516–518
SPSS, 534–535
Scholastic Aptitude Test (SAT), 96, 345
SEM (standard error of the mean), 121–122. *See also* Sampling error
Shendarkar, N. N., 112
Sigma (Σ), as summation notation, 43–45
Significance testing:
null hypothesis in, 5
repeated/related samples *t* test (activity), 299–308
summary of, 344–345
See also ANOVA, two-way
Simple effects analysis for two-way ANOVA, 484–485
Single-sample mean difference, confidence interval for, 252–253
Single sample *t* test, 207–239
alpha levels for, 222–223
conceptual information on, 208–211
independent samples *t vs.*, 315–319
overview, 207–208
repeated/related samples *t* test and, 273–274
SPSS, 223–226
Single sample *t* test, one-tailed:
activity on, 230–232
degrees of freedom and critical regions for, 213–214
effect size for, 216
null and research hypotheses for, 212–213
results interpretation for, 216–217
statistical assumptions for, 211–212
test statistic for, 214–216, 232–233
Single sample *t* test, two-tailed:
activity on, 226–230
degrees of freedom and critical regions for, 220
effect size for, 222
null and research hypotheses for, 219–220
overview, 217–219
results interpretation for, 222
statistical assumptions for, 219
test statistic for, 220–221, 232–233
Single score, *z* scores for, 95–97
Skewed distributions, 16–17, 41–42
Social Loneliness Scale (SLS), 309
Spearman's correlation (r_s):
Pearson's correlation *(r) vs.*, 521–522, 537–539, 560
ranks of scores analyzed by, 532
SPSS, 550–551
table of, 611
SPSS (Statistical Package for the Social Sciences):
central tendency, 47–51
chi-square statistics, 578–583
confidence intervals, 251–252
correlation and regression, 533–536, 550–551
data file, 20–21
frequency distribution tables and graphs, 21–24
hypothesis testing with repeated/related samples *t* test, 295–298
independent ANOVA, 400–410
independent samples *t*, 330–335
one-way ANOVA, 398–400
one-way independent samples ANOVA, 383–385
repeated/related samples *t* test, 284–288
sample means and sample mean difference estimation activity, 261–266
single sample *t* test, 223–226
two-way ANOVA, 458–464
variability, 75–78
Standard deviation:
computation example of, 86–90
distribution of sample means, 118–119
population, 69–72
sample, 74
single sample *t* test, 207
variability measured by, 65–66
Standard error of the mean (SEM), 121–122. *See also* Sampling error
Standard normal curve, *z* scores and, 98–100
Statistical assumptions:
assessing, 146–148
chi-square statistics for goodness-of-fit, 571
chi-square statistics for independence, 573
hypothesis testing with independent samples *t*, 337
hypothesis testing with *z* scores, 166–167
one-tailed independent samples *t*, 326
one-tailed Pearson's correlation, 529
one-tailed repeated/related samples *t* test, 280
one-tailed single sample *t* test, 211–212
one-way independent samples ANOVA, 372–373
sample mean example (one-tailed) of hypothesis testing, 146–148

two-tailed independent samples *t*, 320–321
two-tailed Pearson's correlation, 526–527
two-tailed repeated/related samples *t* test, 274–275
two-tailed single sample *t* test, 219
two-way ANOVA, 445–446
Statistical estimation, 256–257
Statistical Package for the Social Sciences (SPSS). *See* SPSS (Statistical Package for the Social Sciences)
Statistical power:
hypothesis testing with independent samples *t*, 341–344
hypothesis testing with *z* scores (activity), 173–176
overview, 158–159
Statistical power, type I and type II error and:
alpha level change and, 191–193
hypothesis testing for, 185–189
reducing measurement error and, 193
results reporting, 194–195
sample size and, 190–191
treatment effect size and, 189–190
WISE statistical power applet for, 182–185
Statistical significance, 157–158. *See also* Hypothesis testing
Statistics, introduction to, 1–37
causal hypothesis testing, 6–7
discrete *vs.* continuous variables, 12
distributions, shapes of, 15–19
frequency distribution tables, 19
graphing data, 12–15
helping professions and, 4–5
hypothesis testing, effect size, and confidence intervals, 5–6
independent and dependent variables, 8–9
math skills required, 3–4
measurement scales, 9–12
overview, 1–3
populations and samples, 7–8
purpose of, 4
SPSS statistics software, 20–24
Studentized range statistic *(q)* table, 607–608
Sum of the squared deviation scores *(SS)*:
computational formulas for, 68–69, 87
population variability, 68–69
sample variability, 73
standard deviation of the distribution of sample means, 118
Surveys, summed scores from, 11
Symmetrical distributions, 41–42

t distributions, 254–256. *See also* Independent samples *t;* Repeated/related samples *t* test; Single sample *t* test
Test statistic:
chi-square statistics, 588–592
chi-square statistics for goodness-of-fit, 572–573
chi-square statistics for independence, 574–576
choosing, 353–355, 428–430, 616–617
correlation and regression, 560–562
hypothesis testing with *z* scores, 169
one-tailed Pearson's correlation, 530–531
one-tailed repeated/related samples *t* test, 281–283
one-tailed single sample *t* test, 214–216, 232–233
one-way independent samples ANOVA, 374–380
sample mean example (one-tailed) of hypothesis testing, 154–155
two-tailed independent samples *t*, 322–325
two-tailed Pearson's correlation, 528
two-tailed repeated/related samples *t* test, 277–279
two-tailed single sample *t* test, 220–221, 232–233
two-way ANOVA, 452–454, 502–505
Therapeutic Touch activity, 24–27
Treatment effect:
type I and type II error, statistical power and, 189–190
variance affected by, 368, 370
t test. *See* Independent samples *t;* Repeated/related samples *t* test; Single sample *t* test
Tukey's honestly significant difference (HSD) test:
computing, 378–380, 383
independent ANOVA, 409, 427
one-way ANOVA, 400
overview, 395
SPSS, 397–398
Two-factor or two-way analysis of variance (ANOVA). *See* ANOVA, two-way
Two-tailed independent samples *t*:
activity on, 347–353
statistical assumptions, 320–321
null and research hypotheses, 321–322
degrees of freedom and critical region, 322
test statistic, 322–325
effect size, 325
results interpretation, 326
Two-tailed Pearson *(r)* correlation:
critical region, 527–528
effect size, 528–529
null and research hypotheses, 527
overview, 525
results summarization, 529
statistical assumptions, 526–527
table of, 610
test statistic, 528
Two-tailed probabilities *t* table, 602
Two-tailed repeated/related samples *t* test:
degrees of freedom and critical region, 276–277
effect size, 279
hypothesis testing with, 293–295
null and research hypotheses, 275–276
results interpretation, 279–280
statistical assumptions, 274–275
test statistic, 277–279
Two-tailed single sample *t* test:
activity on, 226–230
degrees of freedom and critical regions for, 220
effect size for, 222
null and research hypotheses for, 219–220
overview, 217–219
results interpretation for, 222
statistical assumptions for, 219

test statistic for, 220–221, 232–233
See also Single sample *t* test
Type I and type II error:
alpha level impact on, 222–223
description of, 158–159
hypothesis testing with independent samples *t*, 341–344
hypothesis testing with *z* scores (activity), 173–176
Type I and type II error, statistical power and:
alpha level change and, 191–193
hypothesis testing for, 185–189
reducing measurement error and, 193
results reporting, 194–195
sample size and, 190–191
treatment effect size and, 189–190
WISE statistical power applet for, 182–185

United Nations Development Programme, 560
Unit normal table:
reference, 597–600
z scores and, 101–106
University of Washington, 504
Upper boundary of confidence intervals, 244, 247–248

Variability, 65–94
activity on, 78–90
sample, 72–74
SPSS, 75–78
Variability, population:
deviation score computation, 66–67
overview, 65–66
squaring deviation scores, 67–68
standard deviation computation, 69–72
summing squared deviation scores, 68–69
variance computation, 69
Variability, within- and between-group:
activity on, 411–423
description of, 389–394
Variables:
confounding, 241
continuous *vs.* discrete, 12
independent and dependent, 8–9
statistical assumptions, 146–148
Variance:
computation of, 69, 73
homogeneity of, 146–147, 211, 219
homogeneity of (Levene's test), 332–335
pooled, 328
See also ANOVA, one-way
Violence Risk Appraisal Guide (Harris, Rice, & Quinsey), 5

Welch *t* test, 335n
WISE statistical power applet, 182–185
Within- and between-group variability:
activity on, 411–423
description of, 389–394
Within-subjects samples *t* test, 272
Women's College Coalition, 227, 230
Wunderlich, Carl, 107

x for "cut lines," 97–98

Zero mean difference, 250–251
z scores, 95–112
correlation and, 514–516
distribution of sample means, 122–125
negative, 102–103
positive, 100–102
proportion between two, 103–104
single score, 95–97
standard normal curve and, 98–100
table of, 612
t test *vs.*, 208–209, 232–233
x for "cut lines," 97–98
See also Hypothesis testing with *z* scores
z scores, probabilities and (activity):
body temperature example, 107–110
computing, 106–107
unit normal table, 104–106
z table, 597–600

Means, Variability, and Standardized Scores

	Sample	*Population*
Mean	$M = \frac{\Sigma X}{N}$	$\mu = \frac{\Sigma X}{N}$
Sum of the squared deviations (SS)	$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$	$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$
Standard deviation	$SD = \sqrt{\frac{SS}{N-1}}$	$\sigma = \sqrt{\frac{SS}{N}}$
Variance	$SD^2 = \frac{SS}{N-1}$	$\sigma^2 = \frac{SS}{N}$
***z* for a single score**	$z = \frac{X - M}{SD}$	$z = \frac{X - \mu}{\sigma}$

Test Statistics and Effect Sizes for Comparing Two Means

	Test Statistic	*Effect Size*
***z* for a sample mean**	$z = \frac{M - \mu}{SEM_p}$	$d = \frac{M - \mu}{\sigma}$
	$SEM_p = \frac{\sigma}{\sqrt{N}}$	
Single-sample *t* $df = N - 1$	$t = \frac{M - \mu}{SEM_S}$	$d = \frac{M - \mu}{SD}$
	$SEM_S = \frac{SD}{\sqrt{N}}$	
Related samples *t* $df = N - 1$	$t = \frac{M_D}{SEM_r}$	$d = \frac{M_D}{SD_D}$
	$SEM_r = \frac{SD_D}{\sqrt{n}}$	
Independent samples *t* $df = (n_1 - 1) + (n_2 - 1)$	$t = \frac{(M_1 - M_2)}{SEM_i}$	$d = \frac{M_1 - M_2}{\sqrt{SD_p^2}}$
	$SD_p^2 = \frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{(n_1 - 1) + (n_2 - 1)}$	
	$SEM_i = \sqrt{\frac{SD_p^2}{n_1} + \frac{SD_p^2}{n_2}}$	
Guidelines for interpreting *d*: .2 = small, .5 = medium, .8 = large		

Hypothesis Testing and Effect Sizes for Comparing Three or More Means

One-Way Independent-Measures ANOVA

Source of Variance	*SS*	*df*	*MS*	*F*
Between treatments	$n\left(\Sigma M^2 - \frac{(\Sigma M)^2}{g}\right)$	$g-1$	$\frac{SS_{between}}{df_{between}}$	$\frac{MS_{between}}{MS_{within(error)}}$
Within treatments (error)	$\Sigma SS_{each\ treatment}$	$N-g$	$\frac{SS_{within(error)}}{df_{within(error)}}$	
Total	$SS_{between} + SS_{error}$	$N-1$		

One-Way ANOVA Effect Sizes	
For overall ANOVA: $\eta_p^2 = \frac{SS_{between}}{SS_{between} + SS_{within(error)}}$	η_p^2 .01 = small .06 = medium .14 = large
For pairwise comparisons: $d = \frac{M_1 - M_2}{\sqrt{SD_p^2}}$	*d* .2 = small .5 = medium .8 = large
$SD_p^2 = \frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{(n_1 - 1) + (n_2 - 1)}$	

Two-Way Independent-Measures ANOVA

Source	*SS**	*df*	*MS*	*F*
Between				
Factor A	*	$a - 1$	$\frac{SS_A}{df_A}$	$\frac{MS_A}{MS_{within(error)}}$
Factor B	*	$b - 1$	$\frac{SS_B}{df_B}$	$\frac{MS_B}{MS_{within(error)}}$
A × B interaction	*	$(a-1)(b-1)$	$\frac{SS_{A\times B}}{df_{A\times B}}$	$\frac{MS_{A\times B}}{MS_{within(error)}}$
Within (error)	*	$N-(a)(b)$	$\frac{SS_{within(error)}}{df_{within(error)}}$	
Total	*	$N-1$		

*SS are computed in SPSS in this book.

Hypothesis Testing and Effect Sizes for Correlation and Regression

Correlation and Regression

	Test Statistic	*Effect Size*
Pearson correlation $df = N-2$	$r = \frac{SS_{XY}}{\sqrt{(SS_X)(SS_Y)}}$ $SS_{XY} = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}$ $SS_X = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$ $SS_y = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N}$	r^2 Guidelines for interpreting r^2: .01 = small, .09 = medium, .25 = large
Spearman correlation $df = N-2$	Same as Pearson but use ranked data	r^2 Guidelines for interpreting r^2: .01 = small, .09 = medium, .25 = large
Regression	$\hat{Y} = bx + a$ $b = r\left(\frac{SD_Y}{SD_x}\right)$ $a = M_y - bM_x$	

Two-Way ANOVA Effect Sizes	
For each main effect and interaction: $\eta_p^2 = \frac{SS_{between}}{SS_{between} + SS_{within(error)}}$ For pairwise comparisons: $d = \frac{M_1 - M_2}{\sqrt{SD_p^2}}$ $SD_p^2 = \frac{(n_1-1)SD_1^2 + (n_2-1)SD_2^2}{(n_1-1)+(n_2-1)}$	η_p^2 .01 = small .06 = medium .14 = large d .2 = small .5 = medium .8 = large

Hypothesis Testing and Effect Sizes for Frequency Counts

Chi-Square

	Test Statistic	*Effect Size*
Chi-square goodness of fit $df = (Categories - 1)$	$\chi^2 = \Sigma \frac{(OF - EF)^2}{EF}$	
Chi-square test of independence $df = (Columns - 1) * (Rows - 1)$	$\chi^2 = \Sigma \frac{(OF - EF)^2}{EF}$ $E = \frac{(RT)(CT)}{N}$	For 2 × 2: $\phi = \sqrt{\frac{\chi^2}{N}}$ All others: $\phi' = \sqrt{\frac{\chi^2}{N(df^*)}}$ df^* = (Columns − 1) OR (Rows − 1), whichever is smaller See Chapter 14 for size guidelines

Confidence Intervals

	Confidence Intervals
Estimating a population mean: $df = N - 1$	Upper boundary $= M + (t_{CI})(SEM_S)$ Lower boundary $= M - (t_{CI})(SEM_S)$
Estimating a difference (related samples): $df = N - 1$	Upper boundary $= M_D + (t_{CI})(SEM_r)$ Lower boundary $= M_D - (t_{CI})(SEM_r)$
Estimating a difference (independent samples): $df = (n_1 - 1) + (n_2 - 1)$	Upper boundary $= (M_1 - M_2) + (t_{CI})(SEM_i)$ Lower boundary $= (M_1 - M_2) - (t_{CI})(SEM_i)$
Around Pearson's *r*: $df = N - 2$	Upper boundary $= (z_r) + (z_{CI})\left(\frac{1}{\sqrt{N-3}}\right)$ Lower boundary $= (z_r) - (z_{CI})\left(\frac{1}{\sqrt{N-3}}\right)$

	One-Way ANOVA	*Two-Way ANOVA*	*Chi-Square*
Research Situation	Testing differences between two or more sample means collected from different groups (e.g., freshmen, sophomores, juniors, seniors). Only one IV	Testing differences between four or more sample means where there are two IVs (e.g., Factor 1 Drug: A or B; Factor 2 Gender: Male or Female)	Analyzing frequency counts Use goodness of fit with one grouping variable and test of independence with two
1. Assumptions	-Appropriate measurement -Normality -Independence -Homogeneity of variance	-Appropriate measurement -Normality -Independence -Homogeneity of variance	-Appropriate measurement -Independence -Expected frequency at least five per cell
2. Hypotheses	H_0: All means are equal H_1: At least one mean is different from one other mean	Test for three effects -Main effect of Factor A (e.g., Drug) -Main effect of Factor B (e.g., Gender) -A × B Interaction (e.g., Drug × Gender)	H_0: The frequencies do not deviate from what is expected H_1: The frequencies do deviate from what is expected
3. Critical region	$df_{between} = g - 1$ $df_{within(error)} = N - g$	Between *df* (numerator) -Main effect of Factor A, $df_A = a - 1$ -Main effect of Factor B, $df_B = b - 1$ -A × B Interaction, $df_{A \times B} = (a - 1)(b - 1)$ Within(error) *df* (denominator) $-df_{within(error)} = N - ab$	Goodness of fit $df = (Categories - 1)$ Independence $df = (Columns - 1)\ (Rows - 1)$
4. Test statistic	Overall ANOVA $SS_{between} = n(\sum M^2 - \frac{(\sum M)^2}{g})$ $MS_{between} = \frac{SS_{between}}{df_{between}}$ $SS_{within(error)} = \Sigma SS_{each\ treatment}$ $MS_{within(error)} = \frac{SS_{within(error)}}{df_{within(error)}}$ $F = \frac{MS_{between}}{MS_{within(error)}}$ Post hoc pairwise comparisons $HSD = q\sqrt{\frac{MS_{within(error)}}{n}}$	Compute *F* for Factor A, Factor B, and interaction We use SPSS for these computations and simple effects Each MS is computed as $\frac{SS}{df}$ Each F is computed as $\frac{MS}{MS_{within(error)}}$	$\chi^2 = \Sigma \frac{(OF - EF)^2}{EF}$ For goodness of fit: *EF* for each cell is predicted proportion * Total Frequency For independence: $EF = \frac{(RT)(CT)}{N}$
5. Effect size	Overall ANOVA $\eta_p^2 = \frac{SS_{between}}{SS_{between} + SS_{within(error)}}$ .01, .06, .14 Pairwise comparisons $d = \frac{M_1 - M_2}{\sqrt{SD_p^2}}$	Overall ANOVA $\eta_p^2 = \frac{SS_{between}}{SS_{between} + SS_{within(error)}}$ .01, .06, .14 Pairwise comparisons $d = \frac{M_1 - M_2}{\sqrt{SD_p^2}}$ .2, .5 , .8	For 2 × 2: $\phi = \sqrt{\frac{\chi^2}{N}}$ All others: $\phi' = \sqrt{\frac{\chi^2}{N(df^*)}}$ df^* = (Columns – 1) OR (Rows – 1), whichever is smaller

	One-Way ANOVA	*Two-Way ANOVA*	*Chi-Square*
6. Confidence intervals	CI for mean $M \pm (t_{CI})\left(SD/\sqrt{N}\right)$ CI for each mean difference $(M_1 - M_2) \pm (t_{CI})\left(SD_p/\sqrt{N}\right)$	CI for mean $M \pm (t_{CI})\left(SD/\sqrt{N}\right)$ CI for each mean difference $(M_1 - M_2) \pm (t_{CI})\left(SD_p/\sqrt{N_i}\right)$	
7. Summarize	Report if the overall ANOVA was significant and give the *F:* $F\,(df_{between}, df_{within(error)}) = __, p = __, MSE = __, \eta_p^2 = __.$ *If the ANOVA was significant, report the results of the post hoc tests, indicating which pairwise comparisons are significantly different. Be clear about which means are higher/lower for each pairwise comparison. Give* p *and* d *for each pairwise comparison.*	For both main effects and the interaction, report if the effect was significant and give the *F:* $F\,(df_{between}, df_{within(error)}) = __, p = __, MSE = __, = \eta_p^2 __.$ *Report which marginal means were higher for significant main effects.* *If the interaction was significant, explain the simple effects. Give* p *and* d *for each simple effect.*	The cell frequencies did (or did not) differ from what was expected, $\chi^2\,(df, N = __) = __, p = __, \phi = __.$ *Explain how they differed from what was expected.*
8. **SPSS instructions for significance test**	-Analyze -General linear model -Univariate -Move the IV into the Fixed Factors box and move the DV into the Dependent Variable box. -Click on Options, then check Descriptive Statistics and Estimates of Effect Size -Click on Continue -To obtain post hoc tests, click on the Post Hoc button and then move the IV into the box labeled Post hoc tests for -Select the Tukey checkbox and then Continue -Click on OK to run the ANOVA	-Analyze, General Linear Model, Univariate -Move the DV into the Dependent Variable box -Move both IVs into the Fixed Factors box -Click on Options, Descriptive Statistics, Estimates of Effect Size To obtain confidence intervals: -Move everything that is in the Factor(s) and Factor interactions into the Display Means for box -Click on Compare Main effects -Select LSD(none) from the Confidence Interval Adjustment drop-down box -Click on Continue and then OK To obtain simple effects, you must use syntax. /EMMEANS = TABLES(FactorA) COMPARE ADJ(BONFERRONI) /EMMEANS = TABLES(FactorA*FactorB) COMPARE (FactorB) ADJ(BONFERRONI)	<u>Goodness of fit</u> -Analyze, nonparametric tests, legacy dialogs, chi-square. -Move the variable to the Test Variable List -Click on the Values button and enter the expected frequencies <u>Independence</u> -Analyze, descriptive statistics, crosstabs -Move one variable to the Row(s) box and the other variable to the Column(s) box -Click the Statistics button, select Chi-square and Continue -Click the Cells button, select the Observed and Expected boxes -Click Continue, then click OK

The **Second Edition** takes a unique, active approach to teaching and learning introductory statistics that allows students to discover and correct their misunderstandings as chapters progress rather than as they conclude. Empirically developed, self-correcting activities reinforce and expand on fundamental concepts, targeting and holding students' attention. Based on contemporary memory research, this learner-centered approach leads to better long-term retention through active engagement while generating explanations. Along with carefully placed reading questions, this edition includes learning objectives, realistic research scenarios, practice problems, self-test questions, problem sets, and practice tests to help students become more confident in their ability to perform statistics.

New and Key Features

- **Enhanced activity questions**, with most questions using a fixed-choice format, provide better feedback to students to identify areas of improvement.
- **More instruction on writing results using APA style**, particularly in the t-test and ANOVA chapters, gives readers confidence to convey their data.
- **A decision tree** at the end of the book helps students choose the correct statistical tool for hypothesis testing.
- **Chapter-ending practice tests** provide students with more opportunities for practice.

AN INTRODUCTION TO STATISTICS

EDITION 2



SAGE edge™

Give your students the SAGE edge!

SAGE edge offers a robust online environment featuring an impressive array of free tools and resources for review, study, and further exploration, keeping both instructors and students on the cutting edge of teaching and learning. Learn more at **edge.sagepub.com/carlson2e.**

SAGE www.sagepublishing.com
Los Angeles | London | New Delhi | Singapore | Washington DC | Melbourne

SUSTAINABLE FORESTRY INITIATIVE
Certified Chain of Custody
Promoting Sustainable Forestry
www.sfiprogram.org
SFI-01268
SFI label applies to text stock

Cover Image: ©Iscatel/shutterstock.com

ISBN 978-1-4833-7873-2