Statistical Methods for Psychologists, Volume 1:

Introduction to Statistical Inference and Experimental Design

Douglas G. Bonett
University of California, Santa Cruz

2021

© All Rights Reserved

Version 1.9

Contents

Chapter 1 Statistical Inference **1**

1.1 Introduction 1
1.2 Study Population 1
1.3 Measurement Properties 2
1.4 Population Parameter 3
1.5 Random Samples and Parameter Estimates 4
1.6 Standard Error 6
1.7 Confidence Interval for a Population Mean 6
1.8 Confidence Interval for a Population Total Quantity 8
1.9 Prediction Interval 8
1.10 Choosing a Confidence Level 9
1.11 Hypothesis Testing 10
1.12 Interpreting a p-value 11
1.13 Normal (Gaussian) Curve 12
1.14 Skewness and Kurtosis 14
1.15 Sampling Distribution of $\hat{\mu}$ 15
1.16 Illustration of the Central Limit Theorem 17
1.17 Probability 18
1.18 Uncertainty in Statistical Results 20
1.19 Power of a Hypothesis Test 21
1.20 Target Population 22
1.21 Nonrandom Samples 24
1.22 Assumptions for Confidence Intervals and Tests 25
1.23 Assessing the Normality Assumption 26
1.24 Data Transformations 27
1.25 Distribution-free Methods 28
1.26 Variability Assessment 31
1.27 Relative Variability 32
1.28 Sample Size Requirements for Desired Precision 33
1.29 Sample Size Requirements for Desired Power 34
1.30 Sampling in Two Stages 36
1.31 Specifying Planning Values 37
1.32 Bayesian Approach to Data Analysis 39
1.33 Sampling From a Small Population 43
1.34 Data Screening 44
Key Terms 46
Concept Questions 47
Data Analysis Problems 50
Suggested Readings 52

Chapter 2 Two-group Designs **53**

2.1 Two-group Experimental Designs 53
2.2 Two-group Nonexperimental Designs 54
2.3 Confidence Interval for a Population Mean Difference 55

2.4 Confidence Interval for a Population Standardized Mean Difference 57
2.5 Confidence Interval for a Ratio of Population Means 60
2.6 Prediction Interval for a Score Difference 61
2.7 Directional Two-sided Test 62
2.8 Equivalence Test 63
2.9 Superiority and Noninferiority Tests 64
2.10 Variability Assessment 65
2.11 Assumptions 67
2.12 Distribution-free Methods 70
2.13 Sample Size Requirements for Desired Precision 72
2.14 Sample Size Requirements for Desired Power 75
2.15 Unequal Sample Sizes 77
2.16 Data Visualization 78
2.17 Missing Data 81
2.18 Multiple Response Variables 82
2.19 Exploratory Data Analysis 83
2.20 Ethical Issues 84
Key Terms 87
Concept Questions 88
Data Analysis Problems 91
Suggested Readings 94

Chapter 3 Between-subjects Designs 95

3.1 One-factor Experimental Design 95
3.2 Classification Factors 97
3.3 Linear Contrasts 97
3.4 Standardized Linear Contrasts 99
3.5 Simultaneous Directional Two-sided Tests 101
3.6 Hypothesis Tests for Linear Contrasts 102
3.7 One-way Analysis of Variance 103
3.8 Two-factor Designs 106
3.9 Definition of Effects in Two-factor Designs 108
3.10 Pairwise Main Effects and Simple Main Effects 110
3.11 Main Effect and Simple Effect Linear Contrasts 111
3.12 Two-way Analysis of Variance 112
3.13 Analysis Strategies for Two-factor Designs 114
3.14 Three-factor Designs 115
3.15 Three-way Analysis of Variance 118
3.16 Analysis Strategies for Three-factor Designs 119
3.17 Subpopulation Size Weighting 120
3.18 One-way Random Effects ANOVA 121
3.19 Two-factor Design with a Random Classification Factor 125
3.20 Assumptions 126
3.21 Distribution-free Methods 128
3.22 Multiple Confidence Intervals and Hypothesis Tests in Factorial Designs 129
3.23 Using Prior Information 130
3.24. Sample Size Requirements for Desired Precision 132

3.25 Sample Size Requirements for Desired Power 135
3.26 Data Transformations and Interaction Effects 136
3.27 Data Visualization 138
Key Terms 140
Concept Questions 141
Data Analysis Problems 144
Suggested Readings 147

Chapter 4 Within-subjects Designs 147

4.1 Within-subject Experiments 147
4.2 Confidence Interval for a Population Mean Difference 148
4.3 Confidence Interval for a Population Standardized Mean Difference 149
4.4 Confidence Interval for a Ratio of Population Means 150
4.5 Prediction Interval for a Score Difference 150
4.6 Linear Contrasts 151
4.7 Standardized Linear Contrasts 152
4.8 Directional Two-sided Test 153
4.9 Equivalence Test 154
4.10 Superiority and Noninferiority Tests 155
4.11 One-way Within-subjects Analysis of Variance 155
4.12 Wide and Long Data Formats 157
4.13 Pretest-posttest Designs 157
4.14 Two-factor Within-subjects Experiments 159
4.15 Two-way Within-subjects Analysis of Variance 159
4.16 Two-factor Mixed Designs 162
4.17 Two-way Mixed Analysis of Variance 165
4.18 Counterbalancing 166
4.19 Reliability Designs 170
4.20 Effects of Measurement Error 173
4.21 Assumptions 174
4.22 Missing Data 176
4.23 Distribution-free Methods 177
4.24 Variability Assessment 179
4.25 Using Prior Information 180
4.26 Sample Size Requirements for Desired Precision 181
4.27 Sample Size Requirements for Desired Power 185
4.28 Data Visualization 188
Key Terms 191
Concept Questions 192
Data Analysis Problems 194
Suggested Readings 197

Appendix A. Tables 197
Appendix B. Glossary 203
Appendix C. Answers to Concept Questions 213
Appendix D. Answers to Data Analysis Questions 231
Appendix E. Probability Rules 245

Preface

This book was developed for a one-semester or one-quarter course that introduces psychology graduate students to the basic principles of statistical inference and experimental design. The prerequisites are an undergraduate course in applied statistics and high school algebra. This book (Volume 1) is a prerequisite for Volume 2: *Linear Statistical Models*, Volume 3: *Categorical Data Analysis*, Volume 4: *Multivariate Statistical Methods*, and Volume 5: *Introduction to Meta-analysis*.

This book emphasizes confidence intervals rather than null hypothesis significance testing (NHST). Several useful alternatives to NHST, such as directional two-sided tests, inferiority tests, superiority tests, and equivalence tests, are described in each chapter. Sample size planning is also emphasized. Simple methods are presented to determine the sample size required to obtain a confidence interval with desired precision or a hypothesis test with desired power.

Virtually all of the methods described in this book are based on classical, rather than Bayesian, statistical methods. A primary distinguishing feature of a Bayesian analysis is the use of "subjective probability" to characterize the uncertainty of the observed results in a specific study. Most treatments of classical statistical methods assume a "relative frequency" definition of probability, which has justifiably been criticized by Bayesian statisticians for decades. Subjective probability is used throughout this book (and Volumes 2 to 5) to interpret the classical confidence interval and hypothesis testing results obtained in a specific study. The use of subjective probability to interpret the results of classical statistical methods should appease Bayesian criticisms of the classical method.

Another unique feature of this book (and Volumes 2 to 5) is the use of "design-based inference" rather than "model-based inference". With model-based inference, we assume that the data observed in a given study are generated by some random process from some infinite population. With design-based inference, we assume that the data observed in a given study are produced by a sample of individuals who were randomly selected from a clearly defined and finite "study population" of individuals. The pedagogical benefits of using design-based inference are threefold. First, with design-based inference the population parameter to be examined has a clear and explicit interpretation so then a confidence interval or hypothesis test regarding that parameter will also have a clear and explicit interpretation. Second, the statistical results are naturally applied to the study

population from which the sample was randomly selected. This discourages the common mistake of describing the result of a psychological study as if it applies to everyone, everywhere, and all the time. Third, with design-based inference it is possible to define subjective probability in a very simple way that can describe the uncertainty of confidence interval and hypothesis testing results.

Assumptions and the effects of assumption violations are given for every method. Robust and distribution-free methods are presented as alternatives to the traditional methods in situations where important assumptions of the traditional methods are clearly violated. Many of the recommended statistical methods for psychological research are not currently available in the major statistical packages such as SPSS, SAS, and Stata. These recommended methods are available in the `statpsych` R package which is referred to throughout the book. SPSS is very easy to use and is ideal for data entry, data management, and performing a wide variety of statistical analyses. Unlike SAS and Stata, there is an open source version of SPSS called PSPP. Because of its widespread availability and ease of use, SPSS is referred to throughout the book. Students who are proficient in both SPSS and R will be well equipped to perform the types of statistical analyses that are needed in their graduate programs and professional careers.

This book is much shorter than other graduate level statistics books that have been written for psychology students. The goal here is to present the material in a brief and concise manner with the expectation that the instructor will provide additional examples, demonstrations in the use of SPSS and R, or more statistical theory. To improve readability, citations are not included in the text. Instead, relevant references are given in the `statpsych` documentation, and suggested readings are given at the end of each chapter.

Each chapter ends with a set of key terms, concept questions, and data analysis problems. Appendix B contains a glossary of all key terms, Appendix C contains answers to all concept questions, and Appendix D contains answers to all data analysis problems. Appendix E provides a brief summary of probability rules.

Douglas G. Bonett
Santa Cruz, California

Chapter 1

Statistical Inference

1.1 Introduction

This chapter introduces some basic principles and methods of statistical inference. We begin by defining a population of objects and the process of assigning a numerical score to each object in the population. Several different ways to summarize all of the numerical scores in the population will be presented. Some summaries describe the center of the distribution of scores, and other summaries describe the variability of the scores. The researcher will want to know the value of a summary description for the entire population but will often not have the time or resources to measure every object in the population. In these situations, the researcher could assign numerical scores to a small fraction of the objects in the population. *Statistical inference* methods use the information in the sample of objects to make an inference about specific summary descriptions for the entire population. Although inferences about a population that are based on sample information will not be perfectly precise and must be made with some degree of uncertainty, it is possible to design a study that will nevertheless provide useful information about an entire population. The development of statistical inference methods has been lauded as one of the greatest achievements of the 20th century, and today these methods are routinely used in virtually every field of study.

1.2 Study Population

A *study population* is a clearly defined collection of objects. The objects could be animate (e.g., people, animals, plants) or inanimate (e.g., newspaper articles, TV shows, community gardens). In psychological research, a study population usually consists of a specific group of people such as all UCSC undergraduate students, all preschool children in San Jose, or all Arizona public school teachers. Unless otherwise stated, all of the study populations considered here will consist of a specific group of people.

1.3 Measurement Properties

In addition to specifying the study population of interest, a researcher will specify some attribute to measure. When studying human populations, the attribute of interest could be a specific type of academic ability, a personality trait, a type of psychopathology, some particular behavior (e.g., texting, volunteer work, hours of TV watching), an attitude, an interest, an opinion, or a physiological measure. The measurement of the attribute that the researcher wants to examine is called the *response variable* (or *dependent variable*).

To "measure" an attribute is to assign a numerical value to that person. These measurements can have different properties. A *ratio scale* measurement has the following three properties: 1) a score of 0 represents the complete absence of the attribute being measured, 2) a ratio of any two scores correctly describes the ratio of attribute quantities, and 3) a difference between any two scores correctly describes the difference in attribute quantities. For example, heart rate is a ratio scale measurement because a score of 0 beats per minute (bpm) represents a stopped heart and a heart rate of 100 bpm is twice as fast as a heart rate of 50 bpm. In addition, the difference between two heart rates of 50 and 60 bpm describes the same change in heart rate as the difference between 70 and 80 bpm.

With *interval scale* measurements, a score of 0 does not represent a complete absence of the attribute being measured and a ratio of two scores does not correctly describe the ratio of attribute quantities. However, a difference between any two interval scale scores will correctly describe the difference in attribute quantities. For example, suppose a life satisfaction questionnaire is scored on a 0 to 50 scale with higher scores representing higher levels of life satisfaction. A score of 0 does not indicate a complete absence of life satisfaction nor does a score of, say, 40 represent twice the amount of life satisfaction as a score of 20. However, it is assumed that a difference between two life satisfaction scores correctly describes the difference in life satisfaction so a student who obtained a score of, say, 30 while in college and then obtained a score of 35 after graduation is assumed to have the same level of improvement as a student who scored 20 in college and 25 after graduation. Interval and ratio scale measurements will henceforth be referred to as *quantitative scores*.

With *ordinal scale* measurements, the numbers reflect an ordering of the attribute. For example, ordinal scale measurements of 1, 2, and 3 could be assigned to responses of "disagree", "neutral", or "agree", respectively. In this example, a score of 3 indicates greater agreement than a score of 2, and a score of 2 indicates greater agreement than a score of 1. Ordinal scale measurements lack the important properties of interval scale and ratio scale measurements. Unlike an interval scale measurement, the difference between ordinal scores of 1 and 2 does not necessarily represent the same difference in the attribute as the difference between ordinal scores of 2 and 3 or the difference between ordinal scores of 3 and 4. Unlike a ratio scale measurement, an ordinal scale score of 0 does not represent a complete absence of the attribute.

With *nominal scale* measurements, the numbers are simply names for qualitatively different attributes. For example, a participant's preference for Proposal A, Proposal B, Proposal C, or Proposal D could be coded 1, 2, 3, or 4 for the purpose of some data analysis. But these numbers do not possess any of the properties of ordinal, interval, or ratio scales.

All of the statistical methods in this volume assume that the response variable has been measured on an interval or ratio scale. In Volume 3, the response variable is assumed to have been measured on a nominal or ordinal scale.

1.4 Population Parameter

A *population parameter* is a single unknown numeric value that summarizes the measurements that could have been assigned to all N people in a specific study population. Researchers would like to know the value of a population parameter because this information could be used to make an important decision or advance knowledge in some area of research. The *population mean*, denoted by the Greek letter μ (mu), is a population parameter that is frequently of interest.

Imagine every person in a study population of size N being assigned a quantitative score. A *population mean* (μ) is the average of these N quantitative scores. For example, suppose the study population consists of all 2,450 elementary school teachers in a particular school district. Imagine giving a job burnout questionnaire to all 2,450 teachers. The population mean job burnout score would be

$$\mu = \frac{\sum_{i=1}^{N} y_i}{N} \tag{1.1}$$

where y_i is the quantitative burnout score for the i^{th} teacher. The *summation notation* $\sum_{i=1}^{N} y_i$ is a more compact way of writing $y_1 + y_2 + \cdots + y_N$ and is used in many statistical formulas. The quantitative response variable scores $y_1, y_2 \ldots$ will be referred to as "y scores".

Another important population parameter is the *population standard deviation* which is defined as

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \mu)^2}{N}} \tag{1.2}$$

and describes the variability of the y scores. Note that σ (the Greek letter sigma) cannot be negative. The summation notation $\sum_{i=1}^{N} (y_i - \mu)^2$ is a more compact way of writing $(y_1 - \mu)^2 + (y_2 - \mu)^2 + \cdots + (y_N - \mu)^2$. Note also that if all N scores are identical (i.e., there is no variability), every y_i value would equal μ and then σ would be zero. The squared standard deviation (σ^2) occurs frequently in statistical formulas and is called the *variance*.

1.5 Random Samples and Parameter Estimates

In applications where the study population is large or the cost of measurement is high, the researcher may not have the necessary resources to measure all N people in the study population. In these applications, the researcher could instead take a *random sample* of n people from the study population of N people. In studies where random sampling is used, the study population (also called the *sampling frame*) is defined as the population from which the random sample was obtained. A random sample of size n is selected in such a way that every possible sample of size n will have the same chance of being selected. Computer programs can be used to obtain a random sample of size n from a study population of size N. These programs will randomly generate n integers in the range 1 to N, and the integers are then matched to participant identification numbers. The `random.sample` function in the `statpsych` package will generate a random sample of n integers in the range 1 to N.

A population mean can be estimated from a random sample. The *sample mean*

$$\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{1.3}$$

is an estimate of μ (some statistics texts use $\bar{X}$ to denote the sample mean). A caret (^) is placed over the Greek letter to indicate that it is an estimate of the population parameter and not the actual value of population parameter. The sample mean is an *unbiased* estimate of μ because it is just as likely for $\hat{\mu}$ to be larger than μ as it is to be smaller than μ. The sample mean is also called a *maximum likelihood estimate* of μ because it is the most likely value of μ that would have produced the *y* scores.

In a typical survey study, μ is the mean *y* score for all *N* people in the population. In a *one-group treatment study*, a specific treatment (e.g., a time management training program, an anti-anxiety medication, an SAT practice exam) is given to a random sample of *n* participants. In a one-sample treatment study, μ has an interesting and important interpretation – it is the mean *y* score for all *N* people in the population *if they had all received the treatment*. Remarkably, Equation 1.3 provides an estimate of this population mean even though only a small fraction of the population actually received the treatment.

A standard deviation can be estimated from a random sample. The *sample standard deviation*

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{\mu})^2}{n - 1}} \tag{1.4}$$

is an estimate of σ (some statistics texts use *s* to denote the sample standard deviation), and $\hat{\sigma}^2$ is the *sample variance*. The sample variance is unbiased with $n - 1$ rather than *n* in the denominator.

Of course, we would like to know the exact value of μ but we must settle for an estimate of μ if the study population size is large, the measurement process is costly, or it is not feasible to provide some treatment to everyone in the study population. However, the sample mean by itself can be misleading because $\hat{\mu} - \mu$ will be positive or negative and the direction of the error will be unknown. Furthermore, the difference between $\hat{\mu}$ and μ is due to *sampling error* and the unknown magnitude of the sampling error could be small or large.

1.6 Standard Error

The s*tandard error* of a parameter estimate numerically describes the accuracy of a parameter estimate. A small value for the standard error indicates that the parameter estimate is likely to be similar to the unknown population parameter value (e.g., $\hat{\mu}$ is close to μ), while a large standard error value indicates that the parameter estimate could be very different from the population parameter value.

A standard error of a parameter estimate can be estimated from a random sample. The estimated standard error of $\hat{\mu}$ is given below.

$$SE_{\hat{\mu}} = \sqrt{\frac{\hat{\sigma}^2}{n}} \tag{1.5}$$

From Equation 1.5 it is clear that increasing the sample size (n) will decrease the value of the standard error and hence increase the accuracy of the sample mean. From Equation 1.5, it also can be seen that variability in the quantitative scores affects the accuracy of $\hat{\mu}$ with larger variability leading to less accuracy and smaller variability leading to greater accuracy for a given sample size.

1.7 Confidence Interval for a Population Mean

By using an estimate of μ (Equation 1.3) and its estimated standard error (Equation 1.5), it is possible to say something about the unknown value of μ in the form of a *confidence interval*. A confidence interval is defined by a $100(1-\alpha)\%$ lower limit and an $100(1-\alpha)\%$ upper limit that that is believed to contain an unknown population parameter value with $100(1-\alpha)\%$ confidence ("confidence" is defined in section 1.17).

The $100(1-\alpha)\%$ lower limit for μ is

$$LL = \hat{\mu} - t_{\alpha/2;df} SE_{\hat{\mu}}$$

and the $100(1-\alpha)\%$ upper limit for μ is

$$UL = \hat{\mu} + t_{\alpha/2;df} SE_{\hat{\mu}}$$

where $t_{\alpha/2;df}$ is a two-sided critical *t*-value. The value of $t_{\alpha/2;df}$ can be found in a table of critical *t*-values (see Table 2 in Appendix A) or can be computed using the

`qt` function in R. The symbol *df* refers to *degrees of freedom* and is equal to $n - 1$ in a confidence interval for μ. The value $100(1 - \alpha)\%$ is called the *confidence level*. The width of the confidence interval ($UL - LL$) divided by 2 is called the *margin of error*. Henceforth, the two equations for the lower and upper limits of a confidence interval will be expressed using the following simplified notation.

$$\hat{\mu} \pm t_{\alpha/2;df} SE_{\hat{\mu}} \tag{1.6}$$

SPSS and R will compute Equation 1.6 using a sample of y scores. Equation 1.6 can also be computed from the sample mean and standard deviation using the `ci.mean` function in the `statpsych` package. When reporting a confidence interval for μ, also report $\hat{\mu}$ which is the most likely value of μ given the sample data.

Using a larger sample size will tend to decrease the value of $SE_{\hat{\mu}}$, and a smaller value of $SE_{\hat{\mu}}$ will decrease the width of the confidence interval. A narrow confidence interval is more informative than a wide confidence interval. A confidence interval with a higher confidence level (e.g., 95% vs. 90%) will be more convincing, but increasing the level of confidence will also increase the width of the confidence interval. Section 1.27 describes a simple method to determine the sample size that will give a usefully narrow confidence interval with the desired level of confidence.

Example 1.1. A random sample of 10 second-year college students was obtained from a UCSC directory of about 4,000 second-year students. The 10 students were contacted and asked to complete a Sense of Belonging questionnaire (scored from 15 to 75). The scores for the 10 students are given below.

35 36 44 54 43 36 25 41 40 29

The sample mean, sample variance, and standard error for this sample of 10 students are computed below.

$\hat{\mu}$ = (35 + 36 + … + 29)/10 = 38.3
$\hat{\sigma}^2$ = [(35 – 38.3)² + (36 – 38.3)² + … + (29 – 38.3)²]/(10 – 1) = 66.23
$SE_{\hat{\mu}} = \sqrt{\hat{\sigma}^2/n} = \sqrt{66.23/10} = 2.57$ (*continued*)

For $df = n - 1 = 9$, the critical t-value ($t_{.05/2;9}$) can be found in Table 2 of Appendix A. The critical t-value can also be computed using the R command `qt(1 - .05/2, 9)` which returns 2.26. The 95% lower and upper confidence limits are given below.

$$LL = 38.3 - 2.26(2.57) = 32.5 \qquad UL = 38.3 + 2.26(2.57) = 44.1$$

We can be 95% confident that the mean Sense of Belonging score for the 4,000 UCSC second-year students is between 32.5 and 44.1.

1.8 Confidence Interval for a Population Total Quantity

Recall that the population mean is defined as $\mu = \sum_{i=1}^{N} y_i/N$. In studies where the response variable represents a ratio-scale quantity (e.g., dollar amount, ounces of alcohol consumed, hours of TV viewing per week, etc.) and the exact size of the study population (N) is known, a *population total quantity* defined as $N\mu = \sum_{i=1}^{N} y_i$, could be an interesting value to estimate. An estimate of the total quantity is $N\hat{\mu}$, and a confidence interval for the population total is obtained by simply multiplying the endpoints of Equation 1.6 by N.

Example 1.2. A random sample of 200 students was taken from the UCSC student directory of about 17,500 undergraduate students. Every student in the sample was contacted and asked how much they spent on textbooks in the previous quarter. The 95% confidence interval for μ is [\$365.10, \$496.53], and a 95% confidence interval for the total textbook expenditure for all 17,500 students in one quarter is [\$6,389,250, \$8,689,275].

1.9 Prediction Interval

In studies where a random sample of size n has been obtained, the researcher might want to predict the y score for a single member of the study population. A $100(1-\alpha)\%$ *prediction interval* is a range of plausible y scores for one randomly selected member of the study population and is equal to

$$\hat{\mu} \pm t_{\alpha/2;df}\sqrt{\hat{\sigma}^2 + \frac{\hat{\sigma}^2}{n}} \tag{1.7}$$

where $df = n - 1$. A prediction interval for a single y score will never be narrower, and is often much wider, than a confidence interval for μ. Equation 1.7 can be computed using the `pi.score` function in the `statpsych` package. In a one-group treatment design, the prediction interval is a range of plausible y scores for any

randomly selected person from the study population who would be given the treatment.

Example 1.3. A social isolation questionnaire was given to a random sample of 50 first-year UCSC students. The sample mean is 74.7, the sample variance is 36.0, and the 95% confidence interval for the population mean social isolation score is [73.0, 76.4]. A 95% prediction interval for the social isolation score for any *one* randomly selected first-year student is $74.7 \pm 2.01\sqrt{36.0 + \frac{36.0}{50}}$ = [62.5, 86.9]. We can be 95% confident that for any one randomly selected first-year UCSC student, that student's social isolation score will be between 62.5 and 86.9.

1.10 Choosing a Confidence Level

A larger confidence level is more compelling than a smaller confidence level (e.g., 95% vs. 90%), and a narrower confidence interval width (upper limit minus lower limit) is more informative than a wider confidence interval. A 95% confidence interval represents a good compromise between the level of confidence and the confidence interval width, as shown in Figure 1.1. Notice that the confidence interval width increases almost linearly up to a confidence level of about 95% and then the confidence interval width begins to increase dramatically with increasing confidence. Thus, small increases in the level of confidence beyond 95% produce large increases in the confidence interval width.

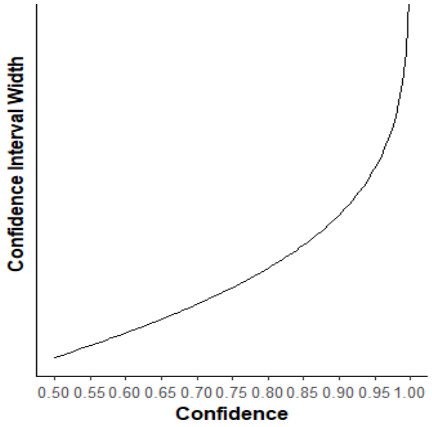


Figure 1.1 Relation between confidence interval width and confidence level

1.11 Hypothesis Testing

In some applications, the researcher simply needs to decide if the population mean is either greater than some value or less than some value. If the population mean is greater than some value, this could provide support for one theory or one course of action. If the population mean is less than some value, this could provide support for another theory or another course of action. This type of decision is called a *directional two-sided test*.

The following notation is used to specify a set of hypotheses regarding μ

H_0: $\mu = h$ H_1: $\mu > h$ H_2: $\mu < h$

where h is some number specified by the researcher and H_0 is called the *null hypothesis*. H_1 and H_2 are called the *alternative hypotheses*. In virtually all applications, H_0 is known to be false because it is extremely unlikely that μ will exactly equal h and the researcher's goal is to decide if H_1 is true or if H_2 is true.

A confidence interval for μ can be used to choose between H_1: $\mu > h$ and H_2: $\mu < h$ using the following rules:

- If the lower limit of a $100(1-\alpha)\%$ confidence interval is greater than h, then H_0 is rejected and H_1: $\mu > h$ is accepted.
- If the upper limit of a $100(1-\alpha)\%$ confidence interval is less than h, then H_0 is rejected and H_2: $\mu < h$ is accepted.
- If the confidence interval includes the value of h, then H_0 cannot be rejected.

A failure to reject H_0 is an *inconclusive result* because we could not decide if $\mu > h$ or $\mu < h$.

A $100(1-\alpha)\%$ confidence interval for μ can also be interpreted as the set of all values of h for which H_0 cannot be rejected. All values of h that are not included in the confidence interval are values for which H_0 would have been rejected at the specified α level. For example, if a 95% confidence interval for μ was [14.2, 18.5], then all tests of H_0: $\mu = h$ will not reject H_0 if h is any value in the range 14.2 to 18.5 but will reject H_0 for any value of h that is less than 14.2 or greater than 18.5.

A *one-sample t-test* can be used to perform a directional two-sided test for a single population mean. The one-sample *t*-test uses a *test statistic* rather than a confidence interval. To test H_0: $\mu = h$ for a specified value of α, the test statistic is

$$t = \frac{\hat{\mu} - h}{SE_{\hat{\mu}}} \tag{1.8}$$

and the following decision rule is used.

accept H_1: $\mu > h$ if $t > t_{\alpha/2;df}$
accept H_2: $\mu < h$ if $t < -t_{\alpha/2;df}$
fail to reject H_0 if $|t| < t_{\alpha/2;df}$

The above decision rule will lead to exactly the same conclusion obtained from a confidence interval. SPSS and R will compute Equation 1.8 using a sample of y scores. The `test.mean` function in the `statpsych` package will compute Equation 1.8 using a sample mean and standard deviation.

Accepting H_1 when H_2 is true or accepting H_2 when H_1 is true is called a *directional error*. When a confidence interval or test statistic is used to select H_1 or H_2, the probability of making a directional error is at most $\alpha/2$ ("probability" is defined in section 1.17).

1.12 Interpreting a *p*-value

SPSS and R will compute a *p-value* that corresponds to the value of *t* in Equation 1.8 (in SPSS output, the *p*-value is labeled "sig"). The *p*-value is simply a transformation of the *t*-value into a scale of 0 to 1. The *p*-value in combination with the sign of *t* can be used to perform a directional two-sided test without referring to a table of critical *t*-values. Specifically, H_0 is rejected if the *p*-value is less than α. If H_0 is rejected, then H_1: $\mu > h$ is accepted if $t > 0$ or H_2: $\mu < h$ is accepted if $t < 0$. If the *p*-value is greater than α, then the results are inconclusive.

It is common practice to report the result of a hypothesis test as "significant" if the *p*-value is less than .05 and "nonsignificant" if the *p*-value is greater than .05. The *p*-value is routinely misinterpreted. Specifically, a "significant" *p*-value does not indicate that an important result has been obtained, and a "nonsignificant" *p*-value does not indicate that the null hypothesis is true. Some researchers report

p-values less than .01 as "highly significant" and *p*-values between .05 and .10 as "marginally significant". This practice is not recommended.

A *p*-value less than α merely indicates that the sample size was large enough to reject the null hypothesis, which was already known to be false. A *p*-value greater than α simply indicates that the results of the hypothesis test are inconclusive. Note also that the probability of making a directional error is determined from the specified value of α and not the *p*-value.

Although *p*-values are reported to three decimal places in most psychology journals, it only matters if the *p*-value is less than α or greater than α. The *t*-value is a function of the sample mean, which contains random sampling error, and so the *p*-value also will contain random sampling error. For example, five random samples of size $n = 50$ were taken from a population where $\mu = 10$ and $\sigma = 5$. The five sample *p*-values for a test of H_0: $\mu = 8$ were .054, .003, .001, .150, and .036, which illustrates the substantial random variability of a *p*-value.

Example 1.4. A random sample of 100 undergraduate social science students completed an advising satisfaction questionnaire that was scored on a 0 to 10 scale. The sample mean is $\hat{\mu} = 7.9$ and the sample standard deviation is 3.05. If the population mean advising satisfaction score is less than 7, more advisers will be hired and all advisers will be given additional training. The 95% confidence interval for μ is [7.3, 8.5]. H_0: $\mu = 7$ can be rejected and H_1: $\mu > 7$ can be accepted. The exact same conclusion could be obtained using the one-sample *t*-test where $t = (7.9 - 7)/\sqrt{3.05^2/100} = 2.95$ and the critical two-sided *t*-value for $\alpha = .05$ is $t_{.05/2;99} = 1.98$. Since 2.95 > 1.98, H_0: $\mu = 7$ can be rejected and H_1: $\mu > 7$ can be accepted. The exact same conclusion could also be obtained if the *p*-value is compared with .05 instead of comparing $t = 2.95$ with $t_{.05/2;99} = 1.98$. Using the `pt` function in R (see Table 2 of Appendix A), the *p*-value for $t = 2.95$ is .004 (SPSS and R will report this *p*-value). This *p*-value is less than .05 and we can reject H_0: $\mu = 7$. And, because the sign of *t* is positive, we accept H_1: $\mu > 7$.

1.13 Normal (Gaussian) Curve

A *histogram* is a graph that visually describes the shape of a distribution of quantitative scores. A histogram is constructed by specifying several equal-length intervals of the quantitative scores and then counting the number of people who have quantitative scores that fall within each interval. An example of a histogram of resilience scores for 495 high school students is shown in Figure 1.2.

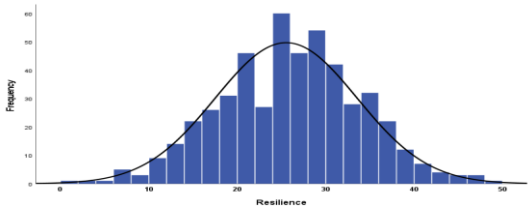


Figure 1.2 Histogram of resilience scores

Scientists discovered decades ago that histograms for many different types of quantitative scores could be closely approximated by a certain type of symmetric bell-shaped curve called a *normal* (or *Gaussian*) *curve*. The histogram of resilience scores in Figure 1.2 includes a graph of a normal curve that, in this example, closely approximates the shape of the histogram.

If a set of quantitative scores is approximately normal, the scores will have the following characteristics:

- about half of the scores are above the mean and about half are below the mean
- about 68% of the scores are within 1 standard deviation of the mean
- about 95% of the scores are within 2 standard deviations of the mean
- almost all (99.7%) of the scores are within 3 standard deviations of the mean

Furthermore, the two points where the normal curve changes from bending down to bending up, called the inflection points, are one standard deviation above and below the mean on the normal curve. A visual inspection of Figure 1.2 suggests that the mean is about 10 and the standard deviation is about 4.

A normal distribution with a mean of 0 and a standard deviation of 1 is called a *standard normal distribution*. If the y scores have an approximate normal distribution, then the *standardized scores* $z_i = (y_i - \mu)/\sigma$ will have an approximate standard normal distribution. The symbol $z_{\alpha/2}$ will be used to denote the point on a standard normal distribution for which $100(1-\alpha)\%$ of the distribution is between $-z_{\alpha/2}$ and $z_{\alpha/2}$. For example, 95% of the standard normal distribution is between $-z_{.05/2}$ and $z_{.05/2}$ where $z_{.05/2} = 1.96$. The equation for the standard normal curve is $f(z_i) = c(e^{-z_i^2/2})$ where $c \approx 0.399$ and $e \approx 2.712$.

1.14 Skewness and Kurtosis

The normal distribution is *symmetric*. In a symmetric distribution, the left half of the distribution is a mirror image of the right half. The asymmetry in a set of quantitative scores can be described using a *coefficient of skewness*. The population coefficient of skewness is equal to $skew(y) = \sum_{i=1}^{N} z_i^3 / N$ where z_i is a standardized score. A skewness coefficient will equal zero if the scores are perfectly symmetric. A skewness coefficient will be positive if the y scores are skewed to the right and will be negative if the y scores are skewed to the left. A positively skewed distribution is shown in Figure 1.3a, and a negatively skewed distribution is shown in Figure 1.3b.

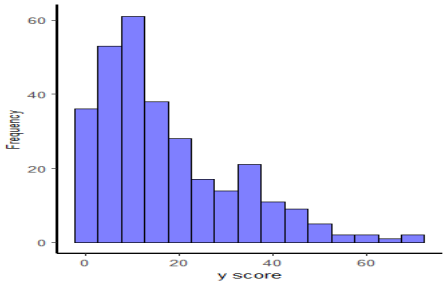


Figure 1.3a Positively skewed distribution

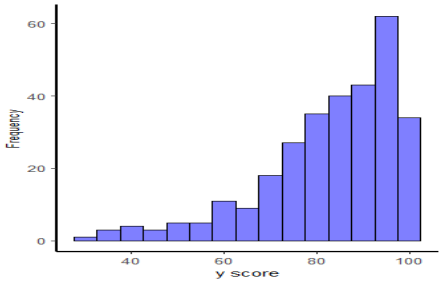


Figure 1.3b Negatively skewed distribution

A distribution of quantitative scores can be nonnormal even if the distribution is symmetric. The *coefficient of kurtosis* describes the degree to which a distribution is more or less peaked than a normal distribution. The kurtosis of a distribution can be described by a coefficient of kurtosis which is equal to 3 in a normal distribution. The population coefficient of kurtosis is equal to $kur(y) = \sum_{i=1}^{N} z_i^4 / N$. SPSS (but not R) subtracts 3 from the kurtosis coefficient so that it will equal 0 in normal distributions. To avoid confusion, a kurtosis coefficient minus 3 is called *excess kurtosis*. *Leptokurtic* distributions have excess kurtosis greater than 0 and are more peaked or have longer tails than a normal distribution. *Platykurtic* distributions have excess kurtosis less than 0 and are less peaked or have shorter tails than a normal distribution. An example of a platykurtic distribution of y scores is shown in Figure 1.4a, and an example of a leptokurtic distribution of y scores is shown in Figure 1.4b. A normal curve is added to each graph for comparison.

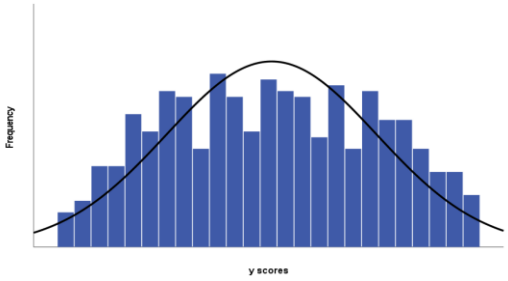


Figure 1.4a Platykurtic distribution

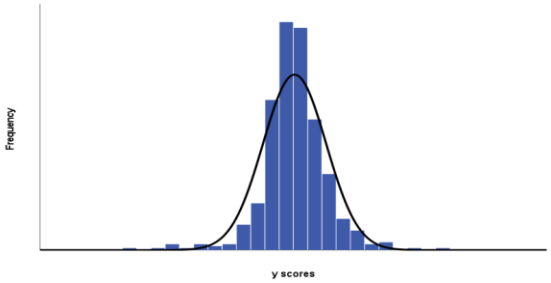


Figure 1.4b Leptokurtic distribution

1.15 Sampling Distribution of $\hat{\mu}$

Consider a study population consisting of N people with y_i representing some quantitative measurement of the i^{th} person. Imagine taking a sample of n people from this study population, recording their y scores, and then computing the sample mean ($\hat{\mu}$). Now imagine doing this for every possible sample of size n. The set of all possible sample means, for samples of size n, is called the *sampling distribution* of the sample mean.

The sampling distribution of $\hat{\mu}$ has three important features:

- The mean of the sampling distribution is equal to the population mean μ
- If the sample size is sufficiently large, the sampling distribution will be closely approximated by a normal distribution regardless of the shape of the distribution of y scores
- The standard deviation of the sampling distribution of the sample means is equal to $\sqrt{\sigma^2/n}\sqrt{(N-n)/(N-1)}$

Because the mean of the sampling distribution of $\hat{\mu}$ is equal to the population mean μ, the sample mean $\hat{\mu}$ is said to be *unbiased*. Unbiased estimates are attractive because they are just as likely to overestimate as underestimate the population parameter.

The standard deviation of the sampling distribution of $\hat{\mu}$ decreases as the sample size increases. If the sample size is large, the sample means in a sampling distribution will have similar values, and because the sample mean is unbiased, they will all tend to be close to the population mean.

In typical applications where n is a small fraction of N, the *finite population correction factor* $\sqrt{(N-n)/(N-1)}$ will be close to 1 and can be ignored. Ignoring the correction factor, the standard deviation of the sampling distribution of $\hat{\mu}$ is $\sqrt{\sigma^2/n}$. Note that the standard error of $\hat{\mu}$ defined in Equation 1.5 is an estimate of the standard deviation of the sampling distribution of $\hat{\mu}$ ignoring the finite population correction. It is remarkable that the standard error of $\hat{\mu}$, which is computed from one random sample, provides an estimate of the standard deviation of the sampling distribution of sample means of all possible samples.

A sampling distribution of $\hat{\mu}$ consists of $N!/[(N-n)!n!]$ values of $\hat{\mu}$, which is an astronomically large number in typical applications (*Note*: $n! = n \times (n-1) \times (n-2) \times \ldots \times 1$; e.g., $4! = 4 \times 3 \times 2 \times 1 = 24$). To concretely illustrate some properties of a sampling distribution, consider a very small population of $N = 5$ people who have quantitative scores of $y_1 = 14$, $y_2 = 13$, $y_3 = 11$, $y_4 = 15$ and $y_5 = 12$ where the population mean is $\mu = (14 + 13 + 11 + 15 + 14)/5 = 13$, and the population variance is $\sigma^2 = [(14-13)^2 + (13-13)^2 + (11-13)^2 + (15-13)^2 + (12-13)^2]/5 = 2$. The standard error of $\hat{\mu}$ for $n = 2$ (not ignoring the finite correction factor) is $\sqrt{\sigma^2/n}\sqrt{(N-n)/(N-1)} = \sqrt{2/2}\sqrt{3/4} = \sqrt{3}/2$. With $N = 5$ and $n = 2$, the sampling distribution of $\hat{\mu}$ consists of only $N!/[(N-n)!n!] = 5!/(3!2!) = 10$ sample means which are shown below.

Sample	Participants	Sample Scores	$\hat{\mu}$
1	1 and 2	14, 13	13.5
2	1 and 3	14, 11	12.5
3	1 and 4	14, 15	14.5
4	1 and 5	14, 12	13.0
5	2 and 3	13, 11	12.0
6	2 and 4	13, 15	14.0
7	2 and 5	13, 12	12.5
8	3 and 4	11, 15	13.0
9	3 and 5	11, 12	11.5
10	4 and 5	15, 12	13.5

The mean of all possible sample means is $(13.5 + 12.5 + \ldots 13.5)/10 = 13$, which is identical to the population mean. Furthermore, the standard deviation of all possible means is $\sqrt{[(13.3-13)^2 + (12.5-13)^2 + \ldots + (13.5-13)^2]/10} = \sqrt{0.75} = \sqrt{3}/2$, which is identical to the standard error of the sample mean.

1.16 Illustration of the Central Limit Theorem

An important theorem in statistical theory is the *central limit theorem*. The central limit theorem states that with a sufficiently large sample size, the shape of the sampling distribution of $\hat{\mu}$ is approximately normal regardless of the shape of the distribution of y scores in the study population. Furthermore, the larger the sample size (up to $N/2$), the more closely the sampling distribution will approximate a normal distribution. Figure 1.5 illustrates a highly nonnormal distribution of y scores in a study population.

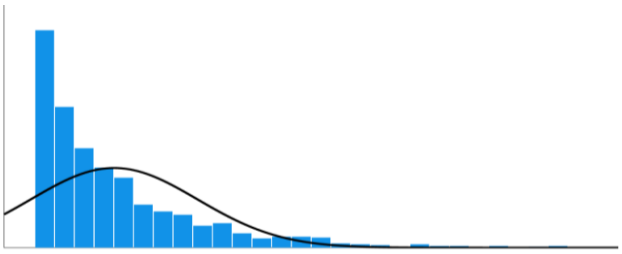


Figure 1.5 Distribution of y scores in a study population

Figures 1.6 and 1.7 illustrate sampling distributions of $\hat{\mu}$ based on samples of $n = 5$, and $n = 30$ (these sampling distributions were approximated by taking 1,000 random samples of a given size from the study population rather than all possible samples). A normal curve is included in each graph for comparison. With samples of size $n = 5$, the sampling distribution of $\hat{\mu}$ is not well approximated by a normal distribution.

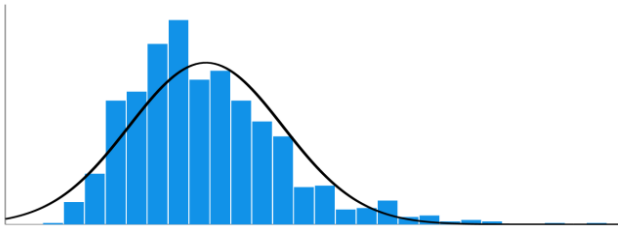


Figure 1.6 Sampling distribution of $\hat{\mu}$ for $n = 5$

With samples of size $n = 30$, the sampling distribution of $\hat{\mu}$ closely approximates a normal distribution even though the y scores in the study population are highly nonnormal.

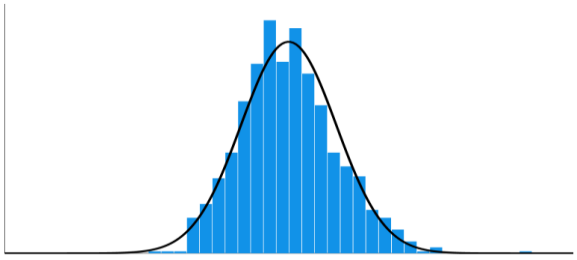


Figure 1.7 Sampling distribution of $\hat{\mu}$ for $n = 30$

In the above example, the distribution of y scores in the study population is extremely nonnormal, but the sampling distribution of $\hat{\mu}$ is closely approximated by a normal distribution with a sample size of 30. If the distribution of y scores in the study population is only moderately nonnormal, the sampling distribution of $\hat{\mu}$ will closely approximate the normal distribution with sample sizes less than 30.

It can be shown that the skewness of a sampling distribution of $\hat{\mu}$ is equal to $skew(y)/\sqrt{n}$, and the excess kurtosis of a sampling distribution of $\hat{\mu}$ is equal to $[kur(y) - 3]/n$. As n increases, the excess kurtosis of a sampling distribution of $\hat{\mu}$ decreases faster than the skewness of a sampling distribution of $\hat{\mu}$. Because of this, skewness is more of a concern than kurtosis when computing a confidence interval for μ or performing a hypothesis test regarding the value of μ.

1.17 Probability

There is an intrinsic amount of uncertainty in all confidence interval and hypothesis testing results. Researchers need to understand and accurately quantify this uncertainly in any reported confidence interval or hypothesis testing result. The uncertainty of a specific outcome can be quantified on a *probability* scale from 0 to 1 where a probability of 0 indicates that some outcome definitely will not occur and a probability of 1 indicates that the event definitely will occur. Two different interpretations of probability, *relative frequency* and *subjective*, are commonly used to describe probability values between 0 and 1. Some basic probability rules are given in Appendix E.

To illustrate the relative frequency approach, imagine an infinitely large population and imagine computing a confidence interval for μ in K different samples from the population. Let f be the number of the K confidence intervals

that include the value of μ. According to the relative frequency definition, the probability that a confidence interval will include μ is equal to f/K as K approaches infinity. The relative frequency definition of probability is useful in theoretical statistics where populations are assumed to be infinitely large and confidence intervals and hypothesis tests are described in terms of imaginary samples from some abstract population. The relative frequency approach is not useful in applied statistics where the populations are finite and it is also necessary to describe the uncertainty of a specific confidence interval or hypothesis test result that has been observed in a single study.

A subjective probability is based on an individual's personal judgment and knowledge about a specific outcome. Unlike the relative frequency interpretation, a subjective interpretation can be used to describe a single outcome. This is important in applied statistics where the researcher conducts one study and must interpret the uncertainty of a confidence interval or hypothesis testing result. *Confidence* can be defined by multiplying a subjective probability by 100%.

When subjective probabilities are assigned to complex phenomena, such as stock prices or weather, people will have differing subjective probabilities about specific outcomes. This lack of consensus is a major criticism of subjective probability. But for very simple phenomena, many individuals can have a consensus opinion about the probability of a specific outcome, and subjective probability is useful in these situations. For example, suppose a jar contains many green and red marbles that are the same size and weight. The marbles are thoroughly mixed and with eyes closed one marble is removed from the jar. Given that the marbles were thoroughly mixed and have the same size and weight, and one was selected with eyes closed, most people would subjectively agree that every marble had the same probability of being selected. This marble example will be more similar to confidence interval and hypothesis testing problems if we also imagine that the marble turns white as soon as it is removed from the jar and that its original color will never be known. Suppose we are told that the proportion of green marbles is .95. In this application, most people would say that they are 95% confident that the selected marble was green. In the following section, subjective probability is used to quantify a researcher's uncertainty regarding confidence interval and hypothesis testing results obtained in a single study.

1.18 Uncertainty in Statistical Results

The subjective probability can be used to interpret a $100(1-\alpha)\%$ confidence interval for μ. If a $100(1-\alpha)\%$ confidence interval for μ was computed from every possible sample of size n in a given study population, we know from statistical theory that about $100(1-\alpha)\%$ of the confidence intervals will include the unknown value of μ. The proportion of all possible samples of a given size for which the $100(1-\alpha)\%$ confidence interval includes the value μ is called the *true coverage probability*.

Consider again the marble example described above. With random sampling, we assume that every possible sample of size n has the same subjective probability of being selected (which is analogous to randomly selecting one marble). We know that each sample will be one of two types: samples where the $100(1-\alpha)\%$ confidence interval contains the value of μ and samples where the $100(1-\alpha)\%$ confidence interval does not include the value of μ (which is analogous to marbles being either green or red). The percentage of all possible samples for which a $100(1-\alpha)\%$ confidence interval includes the value of μ is known to be about $100(1-\alpha)\%$ (which is analogous to knowing to proportion of green marbles). Knowing that a $100(1-\alpha)\%$ confidence interval for μ will include the value of μ in about $100(1-\alpha)\%$ of all possible samples of a given size, and assuming that the one sample the researcher has used to compute the $100(1-\alpha)\%$ confidence interval is a random sample, we can then say that we are $100(1-\alpha)\%$ confident that the computed confidence interval includes the value μ.

For any specified value of α, if a directional two-sided test was performed from every possible sample of size n in a given study population, we know from statistical theory that at most $100\alpha/2\%$ of the hypothesis tests will result in a directional error. The probability of a directional error is close to $\alpha/2$ if μ is close to h but will be less than $\alpha/2$ if μ is not close to h. If we obtain one random sample from the study population and we accept one of the two alternative hypotheses, our subjective probability that we have made a directional error is at most $\alpha/2$. We could also say that we are at least $100(1 - \alpha/2)\%$ confident that we have not made a directional error.

The above subjective interpretations of confidence interval and hypothesis testing results assumed that $100(1-\alpha)\%$ of the confidence intervals from all possible samples of a given size will include the unknown value of μ, and at most $100\alpha/2\%$ of the hypothesis tests from all possible samples of given size will result in a directional error. The conditions required for these claims to be true are described in section 1.22.

1.19 Power of a Hypothesis Test

In hypothesis testing applications, the goal is to reject H_0: $\mu = h$ and then choose either H_1: $\mu > h$ or H_2: $\mu < h$. It is reasonable to assume that H_0: $\mu = h$ is false in any real application because it is extremely unlikely that μ will exactly equal h. The *power* of a hypothesis test is the probability of avoiding an inconclusive result. In a study where the goal is to choose H_1: $\mu > h$ or H_2: $\mu < h$, an inconclusive result would be disappointing. If the power of a hypothesis test is high, then the probability of an inconclusive result will be low. An inconclusive result in a directional two-sided test is also referred to as a *type II error*. The probability of making a type II error is typically denoted by the symbol β, and the power of the hypothesis test is equal to $1 - \beta$.

The power of a directional two-sided test for μ depends on the sample size, the absolute value of $\mu - h$, and the α value. Increasing the sample size will increase the power of the test as illustrated in Figure 1.8 for $\alpha = .05$, $\mu - h = 0.5$, and $\sigma = 1$. Note that increasing the sample size will dramatically increase the power of the hypothesis test up to a point. We typically want the smallest sample size that will produce adequate power. A method for finding the sample size required to achieve desired power is described in Section 1.29.

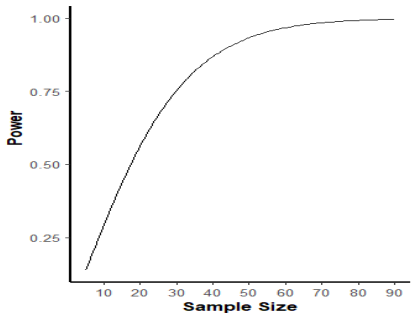


Figure 1.8 Relation between power and sample size

Decreasing α will reduce the probability of a directional error (which is desirable) but will also decrease the power of the directional two-sided test (which is undesirable) as illustrated in Figure 1.9 for $n = 30$, $\mu - h = 0.5$, and $\sigma = 1$. Note that there is little loss in power for reductions in α down to about .10. But the power decreases substantially for α values below .05. This relation between power and α explains why $\alpha = .05$ is a popular choice in psychological research.

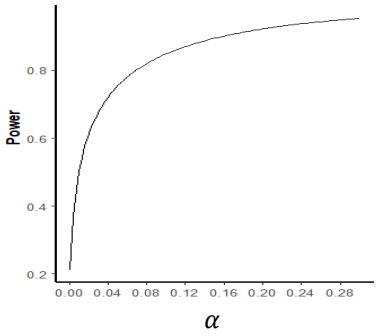


Figure 1.9 Relation between power and α

For a given sample size and α value, Figure 1.10 shows how the power of a directional two-sided test increases as the absolute value of $\mu - h$ increases for $n = 30$, $\alpha = .05$, and $\sigma = 1$. If the value of $|\mu - h|$ is small, a large sample size will be needed to avoid an inconclusive result.

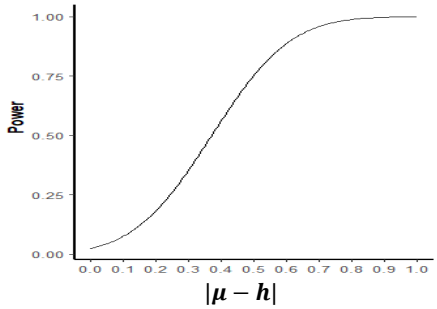


Figure 1.10 Relation between power and $|\mu - h|$

1.20 Target Population

Confidence intervals and hypothesis tests provide information about the study population from which the random sample was taken. In most applications, the study population will be a small subset of some larger and more interesting population called the *target population* (see Figure 1.11). It is important to

remember that the sample mean ($\hat{\mu}$) is an estimate of the study population mean (μ). Furthermore, the target population mean, which will be denoted here as μ^*, is not necessarily similar to the study population mean.

Suppose a researcher obtains a random sample of 100 undergraduate students from a university research participant pool consisting of about 1,000 students. Confidence interval and hypothesis testing results will apply only to those 1,000 undergraduate students, but the researcher is surely more interested in the mean of the response variable for a target population that consists of all undergraduate students.

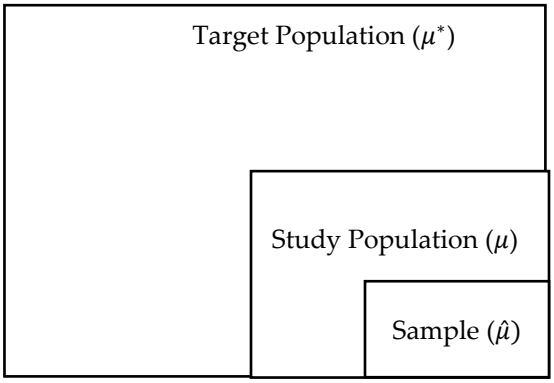


Figure 1.11. The correspondence among the target population, study population, and sample

It might be possible for the researcher to make a persuasive argument that the study population mean should be very similar to the target population mean. If the difference between μ and μ^* is assumed to be trivial, then the confidence interval and hypothesis testing results for μ would then also apply to μ^*. For example, suppose the researcher measured the eye pupil diameter of 100 college students in a small room lit only by a 40-watt light bulb. The researcher could argue that the mean pupil diameter in the study population of 1,000 undergraduate students should be no different than the mean pupil diameter in a target population of all undergraduate students. In this study, it should be easy to convince others that the difference between μ and μ^* is trivial.

Now consider an example where it would be unreasonable to assume that the value of $\mu - \mu^*$ is trivial. Suppose that the researcher instead gave the 100 students a questionnaire to gauge their attitudes about abortion, and also suppose that the university is a Jesuit university. In this study it would not be appropriate to

assume that the confidence interval and hypothesis testing results for μ also apply to a target population of all undergraduate students.

In studies involving sensation, perception, and basic cognitive processes, the value of $\mu - \mu^*$ is typically assumed to be trivial, and researchers in these fields seldom make a distinction between the study and target populations. In contrast, psychologists who study complex human behavior cannot automatically assume that the value of $\mu - \mu^*$ is trivial. In applications where $\mu - \mu^*$ is unlikely to be trivial, the researcher must clearly describe the relevant characteristics of the study population and present the confidence interval and hypothesis testing results in a way that does not give a misleading impression about the generality of the findings.

1.21 Nonrandom Samples

In applications where it is difficult to obtain a random sample from a study population, psychologists will use any easily obtained sample of participants and then "assume" that the sample is random sample from some target population of interest. This assumption is usually very difficult to justify, but in some applications this assumption can be justified. Consider the previous example where a researcher obtained a random sample of 100 college students from a study population of 1,000 students and measured their eye pupil diameters. Instead of taking a random sample, suppose the researcher instead obtained a nonrandom sample of 100 students who were enrolled in an introductory statistics class. The researcher could argue that the nonrandom sample of 100 eye pupil diameters can be thought of as a random sample of the eye pupil diameters that would be obtained in a target population of all young adults. Eye physiology experts would agree with this argument.

It should be noted that a nonrandom sample might be considered a random sample for one response variable but not for other response variables. For example, if the 100 college students in the nonrandom sample described above were given a test to assess their knowledge of basic statistical methods, the test scores for these 100 students is obviously not a random sample of test scores in a target population of all young adults. Examples where a nonrandom sample will yield an interpretable confidence interval or hypothesis testing result are common

in studies of sensation, perception, and basic cognitive processes but are rare in studies involving complex human behavior.

1.22 Assumptions for Confidence Intervals and Tests

The confidence interval and hypothesis test for μ require three assumptions. The importance of obtaining a random sample (the *random sampling assumption*) was made clear in Section 1.18. If the sample is not a random sample from a specific study population and it is not reasonable to assume that the nonrandom sample of scores for a specific response variable could be a random sample of scores from some definable target population, then the confidence interval and hypothesis test results will be uninterpretable.

The *independence assumption* requires the responses from each participant in the sample to be independent of one another. In other words, no participant in the study should influence the responses of any other participant in the study. The *normality assumption* requires the y scores in the study population to have an approximate normal distribution (*Note*: *exact* normality would require an infinitely large study population).

Confidence interval and hypothesis test results will not have the desired interpretation if the independence assumption has been violated. When the independence assumption is violated, the true coverage probability of a $100(1-\alpha)\%$ confidence interval can be far less than $100(1-\alpha)\%$, and the probability of making a directional error in a directional two-sided test can be far greater than $100\alpha/2\%$. Although the consequences of violating the independence assumption are serious, this assumption usually can be easily satisfied by measuring participants one at a time and instructing them not to discuss their responses with any other participants in the study.

One consequence of the central limit theorem is that violating the normality assumption will have little effect on the confidence interval and hypothesis test for μ if the distribution of y scores in the study population is at most moderately skewed and the sample size is not too small ($n > 40$). If the sample size is small and the distribution of y scores in the study population is highly skewed, the true coverage probability of a $100(1-\alpha)\%$ confidence interval will be less than

$100(1-\alpha)\%$, and the probability of making a directional error in a directional two-sided test will be greater than $100\alpha/2\%$.

The `sim.ci.mean` function in the `statpsych` package can be used to examine the performance of the confidence interval for μ with different sample sizes and different nonnormal distributions.

Unlike a confidence interval or a hypothesis test for μ, a prediction interval for a single y score is not protected by the central limit theorem. A prediction interval can have a coverage probability that is lower than the specified level of confidence, even with a large sample size, if the distribution of y scores in the study population is not approximately normal.

1.23 Assessing the Normality Assumption

The confidence interval and hypothesis test for μ assumes that the distribution of the response variable scores in the study population is approximately normal, or that the sample size is large enough that the central limit theorem will provide some assurance that the sampling distribution of the sample mean will be approximately normal. Prediction intervals and some other statistical methods described in subsequent chapters require normality of the response variable scores in the study population, but these methods will not be protected by the central limit theorem. For these statistical methods, the normality assumption must be taken more seriously and researchers must struggle with the fact that the normality assumption can be difficult to assess using only sample data. The shape of the distribution of the y scores in the sample can provide some information about the shape of the population distribution. The estimated skewness coefficient and the estimated kurtosis coefficient can be used to assess the shape of the population distribution. However, estimates of skewness and kurtosis can be inaccurate in small samples.

SPSS and R can be used to compute *bootstrap confidence intervals* for population skewness and kurtosis coefficients. A bootstrap confidence interval is a computationally intensive method that is described in section 2.34 of Volume 2. If a bootstrap confidence interval for population skewness is completely within a -0.5 to 0.5 range, this would be evidence that the population skewness is not

extreme. If a bootstrap confidence interval for population excess kurtosis is completely within a -1 to 1 range, this would be evidence that the population excess kurtosis is not large. However, a large sample size is needed to show that the population skewness is between -.5 and .5 and the population kurtosis is between -1 and 1.

SPSS and R can also be used test the null hypothesis that the population skewness coefficient is zero. If the *p*-value for the test is less than .05, the researcher can conclude that the population scores are not normally distributed and are skewed to the left or to the right according to the sign of the estimated skewness coefficient. The `test.skew` function in the `statpsych` package will compute a *p*-value that is more accurate than the *p*-value computed in most statistical packages. A *p*-value greater than .05 for the test of skewness should not be interpreted as evidence that the normality assumption has been satisfied.

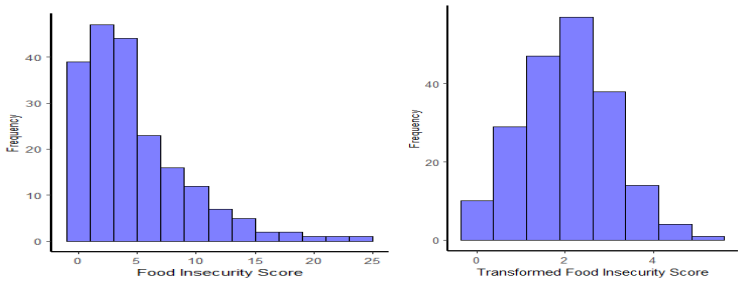
SPSS and R also provide a test of the null hypothesis that the population excess kurtosis coefficient is zero. If the *p*-value for the test is less than .05, the researcher can conclude that the population scores are not normally distributed and are either leptokurtic or platykurtic according to the value of estimated kurtosis coefficient. The `test.kurtosis` function in the `statpsych` package will compute a *p*-value that is more accurate than the *p*-value computed in most statistical packages. A *p*-value greater than .05 for the test of kurtosis should not be interpreted as evidence that the normality assumption has been satisfied.

1.24 Data Transformations

Nonlinear data transformations are sometimes effective in reducing nonnormality in the y scores. When the y score is a frequency count for each participant, such as the number of facts that can be recalled or the number of spelling errors in a writing sample, a square root transformation ($\sqrt{y_i}$) may reduce skewness and kurtosis. When the y score is a time-to-event, such as the time required to solve a problem or a reaction time, a log transformation ($ln(y_i)$) or a reciprocal transformation ($1/y_i$) may reduce skewness and kurtosis. With negatively skewed scores, try transforming $max(y) - y + 1$ where $max(y)$ is the largest y score.

In a *linear* data transformation, each y score is multiplied or divided by a number or a number is added to or subtracted from each y score. A linear data transformation will change the mean and variance of the y scores but will have no effect on skewness or kurtosis. Linear transformations are sometimes used to change the scale of the y scores. For example, if y was measured in seconds then $y/60$ will transform the scores into a scale of minutes.

Example 1.5. A histogram of 200 positively skewed food insecurity scores is shown below (left). A histogram of the square root transformed food insecurity scores (right) is more symmetric and more closely approximates a normal distribution.



Although nonlinear data transformations may reduce nonnormality, the mean of the transformed scores could then be difficult to interpret. However, in some applications the value of μ might be interpretable after a data transformation. For example, if y is measured in squared units, such as the brain surface area showing activity measured in squared centimeters, then $\sqrt{y}$ could be interpreted as the "size" of the activated area. Or if y is the time to respond measured in seconds, then $60/y$ could be interpreted as responses per minute.

1.25 Distribution-free Methods

If the response variable is highly skewed, a population *median* (denoted as θ) could be a more meaningful parameter to estimate than a population mean. The median is useful because it is the value that divides a distribution in half. In skewed distributions, the mean is strongly influenced by a few unusually small or large scores and can give a misleading description of the center of a distribution.

The median also is useful in describing time-to-event scores (e.g., years until divorce, months until next promotion, etc.) which are typically skewed. In a time-to-event study (also called a survival analysis) where participants are studied over

a fixed period of time, some of the participants will not exhibit the event of interest during the study period. We say that the time-to-event scores for these participants are *right censored* because the time-to-event score is some unknown value greater than study period time. If any of the scores are censored, it is not possible to estimate the population mean time-to-event, but the population median time-to-event can be estimated if less than 50% of the scores are censored.

To compute a confidence interval for θ from a random sample n participants with quantitative scores $y_1, y_2, \ldots, y_n$, first rank order the scores from smallest to largest which will be denoted as $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ where $y_{(1)}$ is the smallest score, $y_{(2)}$ is the next smallest score, and $y_{(n)}$ is the largest score. Next, compute $o_1 = (n - z_{\alpha/2}\sqrt{n})/2$ (which is rounded to the nearest integer but not below 1) and $o_2 = n - o_1 + 1$. An approximate $100(1-\alpha)\%$ confidence interval for θ is

$$[y_{(o_1)}, y_{(o_2)}] \tag{1.9}$$

which assumes random sampling, independence among participants, and that tied scores are rare. In a time-to-event study with censored time scores, Equation 1.9 requires $y_{(o_2)}$ to be less than the study period time. Equation 1.9 can be computed using the `ci.median` function in the `statpsych` package. The `sim.median` function in the `statpsych` package can be used to examine the performance of the confidence interval for θ with different sample sizes and different nonnormal distributions. It is instructive to compare the results with those from the `sim.mean` function.

Example 1.6. In *Example* 1.1, the researcher estimated the mean Sense of Belonging score in a study population of about 4,000 UCSC second-year students. The belonging scores in the random sample of 10 students are rank ordered below from smallest to largest

25 29 35 36 36 40 41 43 44 54

where $y_{(1)} = 25$, $y_{(2)} = 29$, … , $y_{(9)} = 44$, $y_{(10)} = 54$. To obtain a 95% confidence interval for θ, compute $o_1 = (10 - 1.96\sqrt{10})/2 = 1.9$ (round to 2) and $o_2 = 10 - o_1 + 1 = 9$. The 95% confidence interval for θ is $[y_{(2)}, y_{(9)}] = [29, 44]$. The researcher can be 95% confident that the median Sense of Belonging score in the study population of 4,000 UCSC second-year students is between 29 and 44.

The sample median, denoted as $\hat{\theta}$, is an estimate of the population median. If n is an odd number, $\hat{\theta}$ is the middle rank ordered score. If n is an even number, $\hat{\theta}$ is the average of the two middle rank ordered scores. For the 10 belonging scores given above, $\hat{\theta}$ = (36 + 40)/2 = 38.

The standard error of $\hat{\theta}$ is approximately

$$SE_{\hat{\theta}} = [y_{(o_2)} - y_{(o_1)}]/(2z_{\alpha*/2}) \tag{1.10}$$

where $y_{(o_1)}$ and $y_{(o_2)}$ are defined in Equation 1.9 with $o_1 = n/2 - \sqrt{n}$, $o_2 = n - o_1 + 1$, and $\alpha^*/2 = .5^n \sum_{i=0}^{o_1-1} \frac{n!}{[i!(n-i)!]}$. The `ci.median` function in the `statpsych` package will compute Equation 1.10.

Equation 1.9 can be used to test the following hypotheses regarding the population median

H_0: $\theta = h$ $\quad$ H_1: $\theta > h$ $\quad$ H_2: $\theta < h$

where h is some number specified by the researcher. Specifically, if the upper limit of the confidence interval is less than h, then H_0 is rejected and H_2 is accepted; if the lower limit of the confidence interval is greater than h, then H_0 is rejected and H_1 is accepted; and if the confidence interval includes h, then H_0 cannot be rejected.

The *sign test* is a distribution-free alternative to the one-sample *t*-test. The sign test is a test of the null hypothesis H_0: $\theta = h$. SPSS and R will compute the *p*-value for the sign test that can be used to decide if H_0 can be rejected. The sign test is preferred to the one-sample *t*-test in applications where the response variable is known to be highly skewed and the sample size is small. The power of the sign test is usually much less than the power of the one-sample *t*-test, but the sign test can have greater power than the *t*-test if the *y* scores are highly leptokurtic.

The null hypothesis H_0: $\theta = h$ for a sign test can also be expressed as H_0: $\pi = .5$ where π is the proportion of people in the study population who have scores greater than h. The results of the sign test can be supplemented with the following approximate $100(1-\alpha)\%$ confidence interval for π

$$\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n+4}} \tag{1.11}$$

where $z_{\alpha/2}$ is a two-sided critical z-value, $\hat{\pi} = (f + 2)/(n + 4)$ and f is the number of participants in the sample with y scores that are greater than h. The `ci.prop1` function in the `statpsych` package can be used to compute Equation 1.11.

1.26 Variability Assessment

The population mean only describes the center of a population of y scores, and it would be a mistake to ignore individual differences and assume that most people have y scores that are similar to the population mean or median. It is important to describe the population variability of the y scores in addition to the population mean or median. The population standard deviation (σ) is a common measure of variability. If σ is small, then most people will have y scores that are similar to the population mean. But if σ is large, then some people will have y scores that are much smaller than μ and others will have y scores that are much larger than μ.

The traditional confidence interval for σ assumes that the population y scores have a normal distribution. This confidence interval will have a coverage probability that can be far less than $1 - \alpha$ if the y scores are leptokurtic, and increasing the sample size will not rectify the problem.

The *mean absolute deviation from the median* (MAD) is an alternative measure of variability that has a simple interpretation and for which a useful confidence interval can be computed. The population MAD is

$$\tau = \frac{\sum_{i=1}^{N} |y_i - \theta|}{N} \tag{1.12}$$

where θ is the population median of the y scores. The summation notation $\sum_{i=1}^{N} |y_i - \theta|$ is a more compact way of writing $|y_1 - \theta| + |y_2 - \theta| + \cdots + |y_N - \theta|$ where the pair of vertical bars represents an absolute value. Thus, the population MAD is simply the average absolute difference between the y scores and the population median. The value of the population MAD is unknown and must be estimated from a random sample. The sample MAD is

$$\hat{\tau} = \frac{\sum_{i=1}^{n} |y_i - \hat{\theta}|}{n} \tag{1.13}$$

where $\hat{\theta}$ is the sample median.

An approximate $100(1-\alpha)\%$ confidence interval for τ is

$$exp[ln(c\hat{\tau}) \pm z_{\alpha/2} SE_{ln(\hat{\tau})}] \quad (1.14)$$

where $c = n/(n-1)$ and $SE_{ln(\hat{\tau})} = \sqrt{[\frac{(\hat{\mu}-\hat{\theta})^2}{\hat{\tau}^2} + \frac{\hat{\sigma}^2}{\hat{\tau}^2} - 1]/n}$. Equation 1.14 assumes the y scores have an approximate normal distribution in the study population, but this assumption is not a concern if $n \geq 30$ and the y scores in the study population are not extremely nonnormal. A less biased estimate of τ is $c\hat{\tau}$. (*Note*: $ln(x)$ is the natural logarithm of x and $exp(x) = e^x$ where $e \approx 2.718$). Equation 1.14 can be computed using the `ci.mad` function in the `statpsych` package.

Example 1.7. A "Feelings of Powerless" questionnaire, scored from 0 to 40, was given to a random sample of 90 students taken from a study population of 1,780 students at a large high school. Scores between 15 and 25 are considered typical, scores above 30 are considered to represent high levels of powerlessness, and scores below 10 are considered to represent low levels of powerlessness. High levels of powerlessness have been associated with susceptibility to conspiracy theories, and low levels of powerlessness have been associated with low susceptibility to conspiracy theories. The 95% confidence interval for the median powerlessness score in the high school student study population is [17.1, 23.5] which is within the typical range. However, the confidence interval for the population MAD is [11.5, 18.8] indicating that there is considerable variability in the powerless scores. The researcher can be 95% confident that the MAD of the powerless scores is between 11.5 and 18.8 points in the study population of 1,780 high school students. Future research will attempt to identify characteristics of those students who exhibit low levels of powerlessness to gain insights that could help develop training programs to reduce susceptibility to conspiracy theories.

1.27 Relative Variability

The standard deviation (σ) and the mean absolute deviation (τ) describe variability in terms of the scale of the y scores. For example, if the y scores are measured in milliseconds, then the values of the standard deviation and mean absolute deviation represent variability in milliseconds. With ratio scale data, the variability of the y scores tends to be larger when the mean is larger, and this complicates the comparison of variability for different demographic groups with differing means. The *coefficient of variation* is a unitless measure of variability that can be used to compare the variability of ratio scale y scores across groups that have different

means. The coefficient of variation is defined as σ/μ. A confidence interval for σ/μ can be obtain using the `ci.cv` function in the `statpsych` package.

The confidence interval for σ/μ assumes approximate normality of the y scores in the population, and the confidence interval can be misleading if the y scores are more than mildly nonnormal. The *coefficient of dispersion* is another unitless measure of variability for ratio scale scores that is defined as τ/θ where θ is the population median. The `ci.cod` function in the `statpsych` package will compute a confidence interval for τ/θ. This confidence interval will perform properly unless the y scores are highly nonnormal.

A third unitless measure of variability for ratio scale scores is the *coefficient of quartile variation,* which is defined as $(\pi_{75} - \pi_{25})/(\pi_{75} + \pi_{25})$ where π_{75} is the 75th percentile (or 3rd quartile) and π_{25} is the 25th percentile (or 1st quartile). The 75th percentile is the y score that is greater than 75% of the scores, and the 25th percentile is the y score that is greater than 25% of the scores. The `ci.cqv` function in the `statpsych` package will compute a distribution-free confidence interval for the population coefficient of quartile variation.

1.28 Sample Size Requirements for Desired Precision

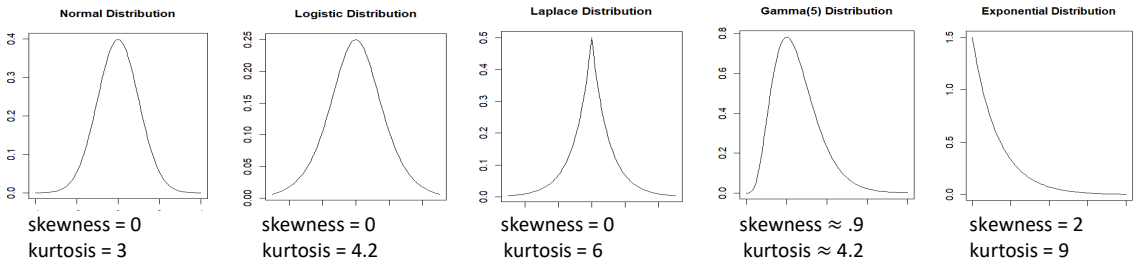
Larger sample sizes give narrower confidence intervals, and it is possible to approximate the sample size that will give the desired width (w) of a confidence interval (i.e., upper limit minus lower limit) for a desired level of confidence. The sample size needed to obtain a $100(1-\alpha)\%$ confidence interval for μ having a desired width of w is approximately

$$n = 4\tilde{\sigma}^2\left(\frac{z_{\alpha/2}}{w}\right)^2 + \frac{z_{\alpha/2}^2}{2} \tag{1.15}$$

where $\tilde{\sigma}^2$ is a *planning value* of the response variable variance and $z_{\alpha/2}$ is a two-sided critical z-value. Equation 1.15 shows that larger sample sizes are needed with: 1) narrower confidence interval widths, 2) greater levels of confidence, and 3) greater variability of the response variable. Equation 1.15 can be computed using the `size.ci.mean` function in the `statpsych` package. Specifying a narrow confidence interval width might require a prohibitively large sample size.

Example 1.8. A researcher wants to estimate the mean empathy score for a population of 4,782 public school teachers. The researcher plans to use an empathy questionnaire (measured on a 1 to 10 scale) that has been used in previous studies. A review of the literature suggests that the variance of the empathy scale is about 6.0. The researcher would like the 95% confidence interval for μ (the mean empathy score for all 4,782 teachers) to have a width of about 1.5. The required sample size is approximately $n = 4(6.0)(1.96/1.5)^2 + 1.92 = 42.9 \approx 43$.

The sample size required to estimate a population median with desired precision depends on the shape of the distribution. The `size.ci.median` function in the `statpsych` package will approximate the required sample size using a variance planning value along with an assumption of the approximate distribution shape. Five distribution assumptions can be specified and are shown below.



In practice, the distribution of y scores in the population will never match any of the above theoretical distributions. If the researcher has any prior information about the approximate skewness and kurtosis of the y scores, this could help select one of the five distribution options. For a given variance planning value, the sample size requirement will be smallest for the Laplace and largest for the Normal distribution. In the absence of any prior information regarding the distribution shape of the y scores, the Normal distribution option will tend to give a conservatively large sample size approximation.

1.29 Sample Size Requirements for Desired Power

The sample size needed in a directional two-sided test of μ with desired power and a specified α value is approximately

$$n = \tilde{\sigma}^2 \frac{(z_{\alpha/2} + z_{\beta})^2}{(\tilde{\mu} - h)^2} + \frac{z_{\alpha/2}^2}{2} \tag{1.16}$$

where $1 - \beta$ is the desired power of the test, $\tilde{\mu}$ is a planning value of the population mean, and z_β is a one-sided critical z-value. The value of $\tilde{\mu} - h$ is the *effect size*.

Equation 1.16 shows that larger sample sizes are needed with smaller values of α, greater desired power, values of μ that are closer to h, and greater variability of the response variable. Equation 1.16 can be computed using the `size.test.mean` function in the `statpsych` package. SPSS can also compute the required sample size for desired power in a one-sample t-test. In applications where the sample size is predetermined, the `power.mean` function will compute the power of the test for the given sample size, α value, and effect size.

Equations 1.15 and 1.16 show that larger values of $\tilde{\sigma}^2$ require a larger sample size. Some researchers prefer to sample from homogeneous study populations (e.g., first year and second year psychology majors) rather than heterogeneous study populations (e.g., working adults) because σ^2 will be smaller in the homogeneous study population and hence the sample size requirement will be smaller. However, hypothesis test and confidence interval results apply to the study population from which the random sample was taken, and the results may have less practical or scientific importance in a homogeneous study population than in a more diverse study population. This tradeoff should be given serious consideration when planning a study.

Example 1.9. A researcher knows that the ACT mathematics scores in a study population of 5,374 first-year college students has a mean of 24.5 and a variance of 8.2. The researcher plans to take a random sample from this study population and then give the sampled students supplementary mathematics training to improve their math skills. The researcher believes that the population mean ACT score would increase from 24.5 to 26.0 if all 5,374 first-year college students were given the supplementary mathematics training. To test H_0: $\mu = 24.5$ for $\alpha = .05$ and a desired power of .90, the required sample size is approximately $n = 8.2(1.96 + 1.28)^2/(26.0 - 24.5)^2 + 1.92 = 40.2 \approx 41$.

The sample size needed to test H_0: $\theta = h$ with desired power and a specified α value using the sign test is approximately

$$n = \frac{[z_{\alpha/2}\sqrt{.25} + z_\beta\sqrt{\tilde{\pi}\,(1-\tilde{\pi})}\,]^2}{4(\tilde{\pi} - .5)^2} + 1/|\tilde{\pi} - .5| \qquad (1.17)$$

where θ is the population median, $\tilde{\pi}$ is a planning value of the proportion of people in the study population who have y scores that are greater than the hypothesized

median (h). The effect size for a sign test is $\tilde{\pi} - .5$. The `size.test.sign` function in the `statpsych` package will compute Equation 1.17.

In sometimes the case that the sample size will not be under the researcher's control but will be determined by practical or monetary considerations. In these situations, we can compute the power of the test for a specified sample size. If the power is unacceptably low, the researcher might decide to abandon the proposed study. The `power.mean` function in the `statpsych` package will compute the power of a one-sample t-test (see Equation 1.8) for a specified sample size.

1.30 Sampling in Two Stages

In applications where sample data can be collected in two stages from the study population, the confidence interval obtained in the first stage can be used to determine how many more participants should be sampled in the second stage. If the $100(1-\alpha)\%$ confidence interval width from a first-stage sample size of n is w_0 and w_0 is larger than the desired width (w), then the number of participants that should be added to the original sample (n^+) to obtain a $100(1-\alpha)\%$ confidence interval width of w is approximately

$$n^+ = \left[\left(\frac{w_0}{w}\right)^2 - 1\right] n. \tag{1.18}$$

The `size.second` function in the `statpsych` package can be used to compute Equation 1.18. Equation 1.18 is general and can be used with any type of confidence interval.

Example 1.10. A researcher computed a 95% confidence interval for a mean gender ideology score (on a 1 to 10 scale) in a population of 1,800 high school students using a random sample of 25 high school students. The width of the confidence interval was 4.38. The results of this study are not adequately informative because the confidence interval is too wide. The researcher would like to obtain a 95% confidence interval for the population mean that has a width of 2.0. To achieve this goal, the number of high school students that should be added to the initial sample is $[(4.38/2.0)^2 - 1]25 = 94.9 \approx 95$ to give a combined sample size of $25 + 95 = 120$.

1.31 Specifying Planning Values

The variance planning value that is used to approximate the sample size required for a confidence interval to have the desired width (Equation 1.15) is a subjective estimate of the *sample* variance that is likely to be observed in the planned study. In comparison, the variance planning value that is used to approximate the sample size required to test a hypothesis with desired power (Equation 1.16) is a subjective estimate of the *population* variance. In practice, the researcher will not know the value of the population variance (needed for Equation 1.16) or what the sample variance will happen to be in a planned study (needed for Equation 1.15).

Subjective variance planning values can be obtained from expert opinion, pilot studies, or a review of published studies that have used the same response variable that will be used in the planned study. If expert opinion or prior variance estimates are not available but if the maximum and minimum possible values of the response variable scale are known, $[(\text{max} - \text{min})/4]^2$ provides a crude planning value of the population variance.

If a prior study has used the same response variable that will be used in the planned study and the prior study has sampled from a population that is similar to the population in the planned study, the sample variance from the prior study can be used to specify the variance planning value in Equations 1.15 and 1.16. However, the sample variance from the prior study will contain sampling error and could understate the value of the population variance required in Equation 1.16. When using Equation 1.16, one option is to compute the following upper one-sided confidence limit for the population variance

$$(n_0 - 1)\hat{\sigma}_0^2 / X_{df;\,\alpha}^2 \tag{1.19}$$

where n_0 is the sample size in the prior study, $\hat{\sigma}_0^2$ is the sample variance in the prior study, $df = n_0 - 1$, and $X_{df;\,\alpha}^2$ is a lower-tail critical chi-square value (see Table 5 of Appendix A). The `ci.var.upper` function in the `statpsych` package will compute Equation 1.19. Unless the sample size in the prior study is large, α in Equation 1.19 could be set to .10 or .20 rather than the traditional .05.

Example 1.11. A food insecurity questionnaire will be given to a random sample of undergraduate students at a large public university. The questionnaire is scored on a 0 to 6 scale with higher scores indicating greater food insecurity. If the population mean insecurity at this university is greater than 3.5, the university will consider creating a food pantry. The food insecurity questionnaire was used in a prior study at a similar public university with a sample of 100 students. The sample variance in the prior study was 1.45, and the upper 90% confidence limit for the population variance was 1.76. For the planned study, the researcher believes that the population mean food insecurity score will be at least 4.0. To test H_0: $\mu = 3.5$ with $\alpha = .05$ and a desired power of .90, the required sample size is approximately $n = 1.76(1.96 + 1.28)^2/(4.0 - 3.5)^2 + 1.92 = 75.8 \approx 76$.

In addition to specifying a variance planning value, Equation 1.16 also requires an effect size planning value ($\tilde{\mu} - h$). Two different approaches can be used to specify the effect size planning value in Equation 1.16. One approach sets the planning value of the mean to its most likely value based on expert opinion or prior research. Another approach sets the value of $|\tilde{\mu} - h|$ to the smallest value that would represent a useful or interesting effect size. This smallest value is referred to as the *minimally interesting effect size*.

Unlike the variance planning value in Equation 1.16, the variance planning value in Equation 1.15 should be a prediction of what the *sample* variance will be in the planned study. If a prior study has used the same response variable and has sampled from a population that is similar to the population in the planned study, the sample variance from the prior study can be used to specify the variance planning value in Equations 1.15. However, if the sample variance from a prior study ($\hat{\sigma}_0^2$) is used as a variance planning value in Equation 1.15, the sample variance in the planned study could be larger than $\hat{\sigma}_0^2$, and then the confidence interval would be wider than desired.

The sample variance from a prior study can be used to predict the value of the sample variance in a planned study using the following one-sided upper prediction limit

$$\hat{\sigma}_0^2 F_{\alpha;n-1;n_0-1} \tag{1.20}$$

where n_0 and $\hat{\sigma}_0^2$ are defined above, n is an initial approximation to the required sample size in the planned study, and $F_{\alpha;n-1;n_0-1}$ is a critical F-value (see Table 4 of Appendix). The initial sample size approximation is obtained using Equation

1.15 with the upper one-sided confidence limit from Equation 1.19. The `pi.var` function in the `statpsych` package can compute Equation 1.20. The result from Equation 1.20 can then be used as a revised variance planning value in Equation 1.15 to obtain a revised sample size requirement. A slightly more accurate sample size approximation for the planned study can be obtained by re-computing Equation 1.20 using the revised sample size requirement and then re-computing Equation 1.15. Unless the sample size in the prior study is large, α in Equation 1.20 could be set to .10 or .20 rather than the traditional .05. The `size.ci.mean.prior` function in the `statpsych` package will compute these multi-step computations.

Example 1.12. A university plans to monitor Zoom fatigue by giving the Zoom Fatigue and Exhaustion (ZFE) questionnaire (scored 1 to 5) to a random sample of undergraduate students every year. A prior study obtained a variance estimate of .71 from a sample of 204 participants. A 90% upper confidence limit for the population variance is .811. Using .811 as the initial variance planning value, the sample size required to estimate the mean ZFE score with 95% confidence and a desired confidence interval width of .4 is approximately $n = 4(.811)(1.96/.4)^2 + 1.92 = 79.8 \approx 80$. A 90% upper prediction limit for the predicted sample variance in the planned study with a sample size of 80 is .895. The revised sample size requirement is approximately $n = 4(.895)(1.96/.4)^2 + 1.92 = 87.9 \approx 88$.

1.32 Bayesian Approach to Data Analysis

There are two fundamentally different approaches to data analysis: the *classical* (or *frequentist*) approach and the *Bayesian* approach. How "probability" is defined is the main difference between these two approaches. Originally, the classical approach relied exclusively on the relative frequency definition of probability (hence the term "frequentist"), while the Bayesian approach has always used the subjective definition. As explained in section 1.17, subjective probability is required to interpret the classical confidence interval or hypothesis testing results that are observed in any single study. Thus, the primary Bayesian criticism of the classical approach can be dismissed.

Classical and Bayesian methods also differ with regards to the use of prior information. In a Bayesian analysis, prior information is expressed in terms of a subjective probability distribution that describes the researcher's personal opinion regarding the plausible values of some population parameter. This distribution is called the *prior distribution*. Then the researcher obtains sample data and the

sample information is combined with the prior distribution to obtain a *posterior distribution*. The posterior distribution is also a subjective probability distribution regarding the plausible values of the population parameter. When estimating μ, it is common to assume that population y scores have a normal distribution and the prior distribution is a normal distribution. With these two assumptions, it can be shown that the posterior distribution will have a normal distribution. The Bayesian estimate of μ is then defined as the mean of the posterior distribution.

Let $\tilde{\mu}$ denote the subjective mean of the prior distribution and let $\sigma_{\tilde{\mu}}$ denote the subjective standard deviation of the prior distribution. The sample mean ($\hat{\mu}$) is defined in Equation 1.3, and its standard error ($SE_{\hat{\mu}}$) is defined in Equation 1.5. The mean of the posterior distribution is denoted as $\tilde{\mu}^*$ and is equal to

$$\tilde{\mu}^* = [(1/\sigma_{\tilde{\mu}}^2)\tilde{\mu} + (1/SE_{\hat{\mu}}^2)\hat{\mu}]/(1/\sigma_{\tilde{\mu}}^2 + 1/SE_{\hat{\mu}}^2) \tag{1.21}$$

which is the Bayesian estimate of μ. The standard deviation of the posterior distribution is denoted as $\sigma_{\tilde{\mu}*}$ and is equal to

$$\sigma_{\tilde{\mu}*} = \sqrt{\frac{1}{1/\sigma_{\tilde{\mu}}^2 + 1/SE_{\hat{\mu}}^2}}\,. \tag{1.22}$$

An approximate 100(1 – α)% *credible interval* for μ is

$$\tilde{\mu}^* \pm z_{\alpha/2}\sigma_{\tilde{\mu}*} \tag{1.23}$$

and is the Bayesian alternative to the confidence interval for μ given in Equation 1.6. The `ci.bayes.normal` function in the `statpsych` package will compute Equations 1.21, 1.22, and 1.23.

The Bayesian estimate of μ is biased in small samples and the bias can be large if $\tilde{\mu}$ is not similar to μ. Furthermore, a 95% credible interval can have a true coverage probability that is far less than .95 unless the sample size is large and $\tilde{\mu}$ is similar to μ. As can be seen in Equations 1.21 and 1.22, if the sample size is large then: 1) $\tilde{\mu}^*$will be similar to $\hat{\mu}$, 2) $\sigma_{\tilde{\mu}*}$ will be similar to $SE_{\hat{\mu}}$, and 3) a 100(1 – α)% credible interval will be similar to a 100(1 – α)% confidence interval.

Example 1.13. In *Example 1.1*, 10 students were contacted and asked to complete a Sense of Belonging questionnaire (scored from 15 to 75). Prior to data collection, suppose "Professor Smith" has an opinion about the plausible values of μ, and Professor Smith's subjective probabilities for these plausible values can be characterized by a normal distribution with a mean of 50 and a standard deviation of 5. After the data are collected, Professor Smith obtains $\hat{\mu}$ = 38.3 and $SE_{\hat{\mu}}$ = 2.57. This sample information alters Professor Smith's opinion about the plausible values of μ. Professor Smith's subjective probabilities about the value of μ can now be characterized by a normal distribution with a mean of $\tilde{\mu}^* = [(1/5^2)50 + (1/2.57^2)38.3]/(1/5^2 + 1/2.57^2) = 40.75$ and a standard deviation of $\sigma_{\tilde{\mu}*} = \sqrt{\frac{1}{1/5^2 + 1/2.57^2}} = 2.29$. The lower and upper limits of Professor Smith's 95% credible interval for μ are *LL* = 40.75 – 1.96(2.29) = 36.3 and *UL* = 40.75 + 1.96(2.29) = 45.2.

The posterior distribution is a subjective probability distribution that depends on the mean and standard deviation of the researcher's subjective prior probability distribution. Different researchers will have different subjective prior probability distributions, which will produce different posterior distributions and different credible intervals. Suppose "Professor Brown" conducted the study in Example 1.13 instead of "Professor Smith" and Professor Brown's opinion about the plausible values of value of μ is characterized by a normal distribution with a mean of 30 and a standard deviation of 3. Professor Brown's subjective posterior probability distribution would have a mean of 34.8 with a standard deviation of 1.95, and the credible interval would be [30.9, 38.6]. Although the credible intervals of Professor Smith and Professor Brown are very different, both are correct because these intervals reflect the differing opinions of the two researchers.

The Bayesian estimate of μ and the credible interval are highly dependent on the mean and standard deviation of the subjective prior distribution. This is a serious limitation in scientific research but much less of an issue in organizational research where the posterior distribution can provide useful information to decision makers within the organization.

The results of a Bayesian analysis can be made less subjective by using a *non-informative* prior distribution where $\sigma_{\tilde{\mu}}$ is set to some large value. With a non-informative prior distribution, the posterior distribution will be determined by the sample data and not the personal opinion of the researcher. The use of a non-informative prior is sometimes referred to as an *objective Bayesian analysis*. As can

be seen in Equations 1.21 and 1.22, if $\sigma_{\tilde{\mu}}$ is many times larger than $SE_{\hat{\mu}}$, then: 1) $\tilde{\mu}^*$ will be similar to $\hat{\mu}$, 2) $\sigma_{\tilde{\mu}*}$ will be similar to $SE_{\hat{\mu}}$, and 3) a $100(1-\alpha)\%$ credible interval will be similar to a $100(1-\alpha)\%$ confidence interval. Using subjective probability to interpret a confidence interval as explained in section 1.17, a $100(1-\alpha)\%$ confidence interval will have the same interpretation as a $100(1-\alpha)\%$ credible interval with a non-informative prior. This equivalence also applies to all of the parameter estimates in Volumes 1, 2, 4, and 5 that have normal sampling distributions and a non-informative normal prior distribution.

Most of the confidence intervals in Chapters 1-4 of this volume, as well as those in Volumes 2-5, can be interpreted as credible intervals with non-informative priors. Instead of computing a Bayesian credible interval with a non-informative normal prior distribution, it is best to compute a classical confidence interval. Most of the $100(1-\alpha)\%$ confidence intervals presented in Volumes 1-5 have been refined over the years and have excellent performance characteristics, while Bayesian credible intervals (like Equation 1.23) can have a true coverage probability that is less than $1-\alpha$ unless the sample size is large.

The directional two-sided hypothesis test (also called a "three-decision rule"), which is described in section 1.11 and will be used in Chapters 2 – 4 (and also in Volumes 2 - 5), is equivalent to a Bayesian hypothesis testing approach where a non-informative prior distribution is assumed for the values of μ. The non-informative prior distribution implies that H_1 is just as likely as H_2 and the decision to choose one of these two alternatives will depend solely on the sample data. The classical directional two-sided hypothesis test has the exact same implication. However, a typical Bayesian hypothesis test would not divide α by 2 so that the probability of making a directional error would be α rather than $\alpha/2$. The use of $\alpha/2$ is recommended here because the two-sided confidence interval results and the hypothesis testing results will then lead to the same conclusion. Also, making a directional error is serious because it can lead to an inappropriate course of action or support an incorrect theory. If all assumptions of the hypothesis test have been satisfied, then the probability of making a directional error will be at most $\alpha/2$ if a two-sided critical *t*-value is used in the decision rule.

Bayesian statisticians have criticized classical methods for ignoring prior information, and until recently, this criticism has been warranted. Classical methods that combine the results from a prior study with the results of a current study are now available and easy to implement. Some examples are given in sections 3.23 and 4.25, and Chapter 4 of Volume 5 is devoted entirely to this topic. Using classical methods to combine a parameter estimate from a prior study with a parameter estimate from a current study has several benefits: 1) the information from a prior study is objective rather than subjective, 2) the replicability of the prior results can be assessed, 3) the confidence interval computed from the combined parameter estimates can be substantially narrower than the confidence intervals in the prior or current studies, and 4) the combined results will generalize to the study populations of the prior study and the current study.

1.33 Sampling From a Small Population

As explained in section 1.15, the standard error of $\hat{\mu}$ is $\sqrt{\sigma^2/n}\sqrt{(N-n)/(N-1)}$ where $\sqrt{(N-n)/(N-1)}$ is the finite population correction factor. In most psychological research applications, n is a small fraction of N and the finite population correction factor can be assumed to equal 1. The confidence interval for μ (Equation 1.6) assumes that n is a small fraction of N, and the sample size formulas in sections 1.28 and 1.29 also assume that n is a small fraction of N.

In some studies, n will not be a small fraction of N. In these situations, the following confidence interval for μ with a finite population correction factor is recommended (use `ci.mean.fpc` in the `statpsych` package). Note that the width of Equation 1.24 approaches 0 as n gets closer to N.

$$\hat{\mu} \pm t_{\alpha/2;df} SE_{\hat{\mu}} \sqrt{\frac{N-n}{N-1}} \tag{1.24}$$

Example 1.10. A random sample of 40 families was obtained from a list of 214 families (obtained from several preschools) where only one parent is a native English speaker. The parents were asked to estimate the number of minutes per day they read to their child. The sample mean is 25.5 and the sample standard deviation is 19.1. A 95% confidence interval for the population mean number of reading minutes per day is [19.4, 31.6]. With a finite population correction, the 95% confidence interval is [20.0, 31.0].

When using Equation 1.24, the sample size (n') required to estimate μ with desired precision is approximately

$$n' = n/(1 + \frac{n-1}{N}) \tag{1.25}$$

where n is given by Equation 1.16.

Example 1.11. A public university has a list of 185 students who have active epilepsy. A random sample of the 185 students will undergo a 45-minute clinical assessment and will be assigned a self-stigma score. The researcher set the variance planning value to 25. The researcher wants the 95% confidence interval for μ, with a finite population correction, to have a width of about 3. The required sample size without a finite population correction is 45 students. With a finite population correction, the required sample size is 37 students.

1.34 Data Screening

Before computing any confidence interval or test statistic, the data should be carefully inspected for errors and suspicious values. Errors should be corrected, but participants with unusually small or large y scores, referred to as *outliers*, need to be assessed on a case-by-case basis to determine a possible cause of the unusual response. An unusual response could be due to an equipment malfunction or a participant who did not follow instructions or was responding carelessly. If a cause of the unusual response can be identified, the researcher could be justified in excluding the participant from the analysis. If a participant is excluded from the analysis, the researcher must disclose this information along with the justification.

Some response variables are the sum or average of responses from multiple questionnaire or test items. In these situations, it is important to also examine the responses to the individual items. When examining item-level responses, look for unusual patterns of responses. For example, if each item is scored on a 1 to 5 scale (e.g., strongly disagree to strongly agree), look for participants who give the same rating to every item. If some items are worded positively and others negatively, participants who give ratings of 4 or 5 on the positively worded items would be expected to give ratings of 1 or 2 on the negatively worded items. If the time to complete the test or questionnaire has also been recorded for each participant, a participant with an unrealistically short completion time could justifiably be excluded from the analysis.

With a small sample, a careful review of every response will not require much effort. If the sample size is large, it will be helpful to let SPSS or R find the minimum and maximum response scores. If the minimum or maximum score is

out of bounds (e.g., the maximum score is 27 on a test where 25 is a perfect score), then make the necessary corrections.

A *box plot* provides information about the shape of the distribution and possible outliers and is a useful visual tool in the data screening process. The box plot of the sample data for Data Analysis Problem 1-1 is shown below.

Figure 1.12. Box plot of dollar donation pledges

The thick line within the box represents the median, the right side of the box represents the 75th percentile (Q3), and the left side of the box represents the 25th percentile (Q1). The 75th and 25th percentiles are the scores that are greater than about 75% and 25% of the scores, respectively.

The interquartile range (IQR) is equal to Q3 – Q1 and contains about 50% of the scores. The two lines to the left and the right of the box are called whiskers. The whisker on the right is equal to Q3 + 1.5(IQR), and the whisker on the left is equal to Q1 – 1.5(IQR). Scores that are greater than the whisker on the right or less than the whisker on the left are represented by small circles and are usually considered to be outliers. Figure 1.12 does not reveal any outliers. If the median is closer to the 25th percentile than the 75th percentile, the data are positively skewed. Alternatively, if the median is closer to the 75th percentile than the 25th percentile, the data are negatively skewed.

A *dot plot* shows the scores for each participant. A dot plot of the sample data for Data Analysis Problem 1-1 is shown below. Participants who have unusually small or large scores should be examined to check for data entry errors or other problems.

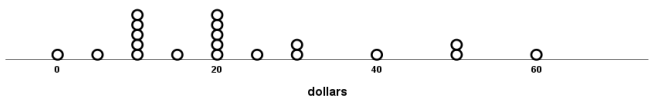


Figure 1.13. Dot plot of dollar donation pledges

Key Terms

study population
response variable
ratio scale
interval scale
ordinal scale
nominal scale
population parameter
population mean
population standard deviation
random sample
sample mean
standard error
confidence interval
prediction interval
directional two-sided hypothesis
one-sample *t*-test
p-value
probability
directional error
power
histogram
normal (Gaussian) curve
coefficient of skewness
coefficient of kurtosis
sampling distribution
central limit theorem
target population
random sampling assumption
independence assumption
normality assumption
data transformations
population median
sign test
relative variability
planning value
box plot
dot plot

Concept Questions

1. Explain in words how to compute a sample variance.

2. Increasing the sample size will have what effect on the width of the confidence interval?

3. Increasing the level of confidence will have what effect on the width of the confidence interval?

4. Increasing the sample size will have what effect on the value of the standard error?

5. Explain in words how a confidence interval for μ can be used to test a direction two-sided hypothesis where H_0: $\mu = 100$.

6. Increasing the sample size will have what effect on the power of a test?

7. Explain in words how to compute a 95% confidence interval for a population mean after you have computed the sample mean and variance in a sample of $n = 20$.

8. When planning a future study to estimate μ, what is the effect of decreasing the desired confidence interval width on the sample size requirement?

9. When planning a future study to estimate μ, what is the effect of increasing the desired level of confidence on the sample size requirement?

10. When planning a future study to estimate μ, what is the effect of using $\tilde{\sigma}^2 = 50$ rather than $\tilde{\sigma}^2 = 75$ on the sample size requirement?

11. When planning a future study to test H_0: $\mu = h$, how does the desired power affect the sample size requirement?

12. When planning a future study to test H_0: $\mu = h$, how does the α value affect the sample size requirement?

13. Why are narrow confidence intervals desirable?

14. Why are confidence interval results more informative than a directional two-sided test?

15. Describe the standard error in the context of a sampling distribution.

16. Why is the 95% level of confidence a popular choice?

17. What are the assumptions of a confidence interval or test for μ, and what are the effects of violating those assumptions?

18. What are some ways to obtain a planning value for σ?

19. Explain why a confidence interval might be uninterpretable if a nonrandom sample is used.

20. Suppose a 95% confidence interval for μ was [1.5, 4.8]. In a test of the null hypothesis H_0: $\mu = h$, for which of the following values of h would H_0 be rejected: 1.20, 3.30, 4.1, or 5.0?

21. How does the size of the study population effect the sample size requirement when n/N is small?

22. Why are data transformations sometimes used?

23. Is it appropriate to declare a null hypothesis such as H_0: $\mu = h$ to be true if the results are "nonsignificant"? Why?

24. When testing a null hypothesis such as H_0: $\mu = h$, what can be said about a "significant" result?

25. What is the relation between the size of the p-value and the sample size?

26. When would a confidence interval for a population median be preferred to a confidence interval for a population mean?

27. Why is it important to assess the variability of the response variable and not just the mean or median of the response variable?

28. What are the implications of the central limit theorem in terms of how a confidence interval for μ will perform?

29. About 1.8 million people visit the Monterey Bay Aquarium each year. Exit interviews were conducted in 2019 for a random sample of 2,417 visitors. One question asked respondents to rate the quality of the educational experience on a 1 to 10 scale. The 95% confidence interval for μ was [8.78, 8.89]. Interpret this result.

30. Explain why each of the following interpretations of [8.78, 8.89] in the above example is incorrect or inadequate.

a) We are 95% confident that the 1.8 million visitor ratings are between 8.78 and 8.89.
b) We are 95% confident that a typical visitor rating will be between 8.78 and 8.89.
c) We are 95% confident that the mean rating of the 2,417 visitors is between 8.78 and 8.89.
d) We are 95% confident that μ is between 8.78 and 8.89.
e) We are 95% confident that 95% of the scores are between 8.78 and 8.89.

Data Analysis Problems

1-1. A researcher wants to estimate the average dollar amount that a Watsonville resident would donate to a local homeless shelter. Twenty residents were randomly selected from the Watsonville directory of residential addresses which contains about 13,800 addresses. The 20 randomly selected residents were contacted by mail and asked how much they would be willing to donate. Their responses (in dollars) are given below.

```
30 20 15 10 10 60 20 25 20 30 10 5 50 40 20 10 10 0 20 50
```

a) Describe the study population.

b) Describe μ in the context of this study.

c) A volunteer group has agreed to contact all 13,800 addresses and ask for a donation if there is strong evidence to suggest that $\mu > 15$. Use SPSS or R to test H_0: $\mu = 15$ with $\alpha = .05$. Report *t, df,* and *p*-value in APA style and state your conclusion.

d) Use R or SPSS to compute a 95% confidence interval for μ and interpret the result.

e) Hand compute a 95% confidence interval for the total dollar amount that the volunteer group should receive if they contact all 13,800 addresses.

f) Use the `ci.median` function in the `statpsych` package to compute a 95% confidence interval for the population median donation amount and interpret the result.

g) A similar study is being planned for the city of San Luis Obispo. How many addresses should be contacted to obtain a 95% confidence interval for μ that has a width of $10.00? Use the sample variance (the squared standard deviation) from the Watsonville study as the planning value for the population variance.

1-2. A national survey suggests that adults sleep an average of 6.8 hours during workdays. Adults who routinely get less than 6 hours of sleep have an increased risk of depression, memory loss, illness, and reduced job performance. A researcher suspects that women living in low-income households may sleep less than the national average. Fifteen women were randomly selected from a social services directory of about 4,000 low-income women in Oakland, CA. Each woman was asked to estimate the number of hours she slept on a typical workday. The number of hours reported by each woman is given below.

```
5.5 5.0 6.5 7.0 4.5 6.0 5.0 7.5 5.0 6.0 8.0 5.0 6.5 5.5 7.0
```

a) Describe the study population.

b) Describe μ in the context of this study.

c) Use SPSS or R to test H_0: $\mu = 6.8$ with $\alpha = .05$. Report *t, df,* and *p*-value in APA style and state your conclusion.

d) Use SPSS or R to compute a 95% confidence interval for μ and interpret the result.

e) Use the `ci.mad` function in the `statpsych` package to compute a 95% confidence interval for the population mean absolute deviation and interpret the result.

f) A similar study is being planned for Dallas, TX. How many low-income women should be sampled to obtain a 95% confidence interval for μ that has a width of 0.5? Use the sample variance (squared standard deviation) from the Oakland study as the planning value for the population variance.

1-3. A critical thinking questionnaire was given to a random sample of 250 adults taken from a listing of about 94,800 Santa Cruz County households. Possible scores range from 0 to 50 with higher scores reflecting higher levels of critical thinking. The sample mean is 41.2 and the sample standard deviation is 2.37.

a) Describe the study population.

b) Describe μ in the context of this study.

c) Use the `ci.mean` function in the `statpsych` package to compute a 95% confidence interval for μ and interpret the result.

d) Use `pi.score` function in the `statpsych` package to compute a 95% prediction interval for the critical thinking score of one adult and interpret the result.

Suggested Readings

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge. **Chapters 1, 2, 3, and 4**

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods, 5*, 411–414. https://psycnet.apa.org/doi/10.1037/1082-989X.5.4.411

Howell, D. C. (2013). *Statistical methods for psychology (8th ed.)* Thompson & Wadsworth. **Sections 2.7, 2.8, 2.9, and Chapter 3**

Snedecor, G. W. & Cochran, W. G. (1989). *Statistical methods (8th ed.)*. Iowa State Press. **Chapters 4 and 5**

Chapter 2

Two-group Designs

2.1 Two-group Experimental Designs

The goal of most research is to assess a possible causal relation between the response variable and another variable called the *independent variable*. In a two-group experimental design, the independent variable has two *levels* that represent two different treatment conditions. In a two-group experimental design, a random sample of n participants is selected from a study population and the random sample is then randomly divided into two groups of sizes n_1 and n_2. The group sample sizes are usually, but are not necessarily, equal. Each group then receives one of the two treatment conditions with participants treated identically within each group. Following treatment, a measurement of the response variable is obtained for each participant.

In a two-group experiment with a quantitative response variable, a population mean could be estimated from each group. In an experimental design, the population means have interesting and important interpretations: μ_1 is the population mean of the response variable assuming everyone in the specified study population had received level 1 of the independent variable (Treatment 1), and μ_2 is the population mean of the response variable assuming everyone in the specified study population had instead received level 2 of the independent variable (Treatment 2). The difference in population means $(\mu_1 - \mu_2)$ for the two treatment conditions is called an *effect size* and describes the strength of the relation between the response variable and the independent variable. If one of the levels of the independent variable is a *control* condition where participants do not receive any treatment, then $\mu_1 - \mu_2$ is sometimes referred to as the *average treatment effect*. Other measures of effect size for a two-group study are described in sections 2.4, 2.5, and 2.12.

Three conditions must be satisfied to demonstrate a causal relation between a response variable and an independent variable. First, there must be a relation

between the response variable and the independent variable (i.e., the effect size must be nonzero). Second, there can exist no other variable that has a causal effect on the response variable and the independent variable. A variable that has a causal effect on the independent variable and the response variable is called a *confounding variable*. Third, the nonzero effect occurs *after* exposure to the different levels of the independent variable.

In an experiment, a nonzero effect size is evidence that the independent variable has a causal effect on the response variable because all three conditions required for a causal relation will have been satisfied: 1) a nonzero effect size implies a relation between the response variable and the independent variable, 2) no variable can be related to the independent variable because the participants were randomly assigned to the levels of the independent variable (see section 2.17 for exceptions), and 3) the nonzero effect will have occurred after exposure to the different levels of the independent variable.

2.2 Two-group Nonexperimental Designs

The benefits of an experiment can be more fully appreciated by considering a two-group *nonexperimental design* where participants are *classified* into two groups according to some preexisting characteristic (e.g., Democrat/Republican, male/female, sophomore/junior, etc.) rather than being randomly assigned into the treatment conditions. Some of the designs in Chapter 3 use a combination of classification and random assignment.

In nonexperimental designs, the effect size describes the strength of a relation between the response variable and independent variable, but this relation cannot be interpreted as a causal relation. In a nonexperimental design, a relation between the independent variable and response variable could be a consequence of one or more confounding variables that have a causal effect on both the response variable and the independent variable. For example, suppose a study finds that the mean GPA for students living off campus is higher than the mean GPA of students living on campus. Students living on campus and off campus might differ in terms of their age, maturity, or study habits. It is possible that one or more of these variables has a causal effect on GPA and where the student chooses to live.

If a relation between an independent variable and a response variable is detected in a nonexperimental design, there is no reason to expect that a person's score on the response variable would change if that person could be placed into the other level of the independent variable. For example, students who are currently living on campus should not expect to see their GPAs increase if they decide to move off campus.

In a nonexperimental design, the parameters also have a different interpretation. Specifically, μ_1 is the population mean of the response variable for all people in one study population who belong to one category of the independent variable (e.g., male, Democrat, sophomore), and μ_2 is the population mean of the response variable for all people in a second study population who belong to the other category of the independent variable (e.g., female, Republican, junior). The members of the study population within each category are referred to as *subpopulations*. The subtle but important parameter interpretation differences in experimental and nonexperimental designs will affect how the researcher describes the results of a confidence interval or hypothesis test.

Two types of sampling can be used in a nonexperimental design. With simple random sampling described in Chapter 1, a random sample of size n is selected from the study population and the participants are then classified into two groups according to some preexisting characteristic. With simple random sampling, the group sample sizes (n_1 and n_2) will not be known in advance and it is possible that one group size could be very small. An alternative sampling method is *stratified random sampling* where the study population is stratified into two subpopulations of sizes N_1 and N_2. Then a random sample of size n_1 is taken from subpopulation 1 and a random sample of size n_2 is taken from subpopulation 2. With stratified random sampling the values of n_1 and n_2 are determined by the researcher.

2.3 Confidence Interval for a Population Mean Difference

A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\hat{\mu}_1 - \hat{\mu}_2 \pm t_{\alpha/2;df} SE_{\hat{\mu}_1 - \hat{\mu}_2} \tag{2.1}$$

where $t_{\alpha/2;df}$ is a critical t-value, $df = \left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2 / [\frac{\hat{\sigma}_1^4}{n_1^2(n_1-1)} + \frac{\hat{\sigma}_2^4}{n_2^2(n_2-1)}]$, and $SE_{\hat{\mu}_1 - \hat{\mu}_2} = \sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}$ is a separate-variance standard error. If the sample sizes are approximately equal, then the separate-variance standard error could be replaced with a pooled-variance standard error $SE_{\hat{\mu}_1 - \hat{\mu}_2} = \sqrt{\frac{\hat{\sigma}_p^2}{n_1} + \frac{\hat{\sigma}_p^2}{n_2}}$ where $df = n_1 + n_2 - 2$, and $\hat{\sigma}_p^2 = [(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2]/df$. The pooled-variance standard error assumes equal population variances and should be used only if the sample sizes are approximately equal.

SPSS and R will compute both the separate-variance and pooled-variance versions of Equation 2.1 using the y scores from each group. The `ci.mean2` function in the `statpsych` package will compute both versions of Equation 2.1 using the sample means and standard deviations from each group.

Example 2.1. 80 first-year college students were randomly selected from a population of about 4,000 new students who are required to take a 2-week orientation course. The 80 students were randomly divided into two groups of equal size. Group 1 was given a set of cultural sensitivity readings, and group 2 received a cultural sensitivity video presentation. Two months later, all 80 students completed a cultural sensitivity behavior checklist, scored from 0 to 25, where they indicated if they had or had not engaged in any of 25 different culturally insensitive behaviors during the last two months (lower scores are better). The sample means and standard deviations are given below.

Group 1	Group 2
$\hat{\mu}_1 = 19.4$	$\hat{\mu}_2 = 11.3$
$\hat{\sigma}_1 = 2.7$	$\hat{\sigma}_2 = 2.1$

The 95% confidence interval for $\mu_1 - \mu_2$ is

$$19.4 - 11.3 \pm t_{.05/2;df}\sqrt{\frac{2.7^2}{40} + \frac{2.1^2}{40}} = [7.02, 9.18]$$

where $df = (\frac{2.7^2}{40} + \frac{2.1^2}{40})^2 / [\frac{2.7^4}{40^2(39)} + \frac{2.1^4}{40^2(39)}] = 73.5$ and $t_{.05/2;73.5} = 2.00$. The researcher is 95% confident that in the study population of 4,000 new college students, the mean score on the cultural sensitivity checklist would be 7.02 to 9.18 lower if they had all received the video presentation rather than a reading list.

2.4 Confidence Interval for a Population Standardized Mean Difference

If the response variable is measured on a familiar scale (e.g., seconds, inches, number of questions answered correctly) the confidence interval for $\mu_1 - \mu_2$ will provide a useful description of the magnitude and importance of the effect of the independent variable. However, many response variables used by psychologists are measured using scales that may not be familiar to other researchers. In these cases, the confidence interval for $\mu_1 - \mu_2$ may not convey easily interpretable information about the magnitude and importance of the effect of the independent variable. For example, suppose a researcher compared two different counseling approaches for test anxiety. Following counseling, the researcher administered a test anxiety questionnaire to all student participants and obtained a 95% confidence interval for $\mu_1 - \mu_2$ equal to [4.23, 7.37]. Is this an important effect? The importance of this result is impossible to describe in the absence of information about the test anxiety scale. To understand the psychological importance of a confidence interval for $\mu_1 - \mu_2$, it would be necessary to at least know the mean and standard deviation of the response variable. In addition, it would be helpful to understand the psychological meaning of different response variable values.

In applications where the psychological meaning of $\mu_1 - \mu_2$ might be unclear to the intended audience, it will be helpful to report a confidence interval for a *standardized mean difference*. One type of population standardized mean difference that is appropriate for experimental designs is

$$\delta = \frac{(\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}} \tag{2.2}$$

and the denominator of Equation 2.2 is called the *standardizer*.

An approximate $100(1-\alpha)\%$ confidence interval for δ is

$$\hat{\delta} \pm z_{\alpha/2} SE_{\hat{\delta}} \tag{2.3}$$

where $\hat{\delta} = (\hat{\mu}_1 - \hat{\mu}_2)/\hat{\sigma}$, $SE_{\hat{\delta}} = \sqrt{\frac{\hat{\delta}^2\left(\frac{\hat{\sigma}_1^4}{n_1 - 1} + \frac{\hat{\sigma}_2^4}{n_2 - 1}\right)}{8\hat{\sigma}^4} + \frac{\hat{\sigma}_1^2}{\hat{\sigma}^2(n_1 - 1)} + \frac{\hat{\sigma}_2^2}{\hat{\sigma}^2(n_2 - 1)}}$, and $\hat{\sigma} = \sqrt{(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2}$.

The following population standardized mean difference uses a weighted variance standardizer and is appropriate for nonexperimental designs

$$\delta' = \frac{(\mu_1 - \mu_2)}{\sqrt{\pi_1\sigma_1^2 + \pi_2\sigma_2^2}} \tag{2.4}$$

where π_1 is the proportion of the study population members that belong to subpopulation 1 and π_2 is the proportion of the study population members that belong to subpopulation 2.

An approximate $100(1-\alpha)\%$ confidence interval for δ' is

$$\widehat{\delta'} \pm z_{\alpha/2}SE_{\widehat{\delta}} \tag{2.5}$$

where $\widehat{\delta'} = (\hat{\mu}_1 - \hat{\mu}_2)/\hat{\sigma}_p$ and $SE_{\widehat{\delta}} = \sqrt{\frac{\widehat{\delta}'^2\left(\frac{1}{n_1-1} + \frac{1}{n_2-1}\right)}{8} + \frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_p^2 n_1} + \frac{\widehat{\sigma}_2^2}{\widehat{\sigma}_p^2 n_2}}$. The estimate $\widehat{\delta'}$ is also known as Cohen's d. SPSS will compute a confidence interval for δ' using a method that assumes equal population variances.

Equations 2.3 and 2.5 do not assume equal population variances. Equation 2.5 assumes simple random sampling. If the population variances are approximately equal, Equation 2.5 can also be used in nonexperimental designs with stratified random sampling or experimental designs.

A third type of standardized mean difference, $\delta'' = (\mu_1 - \mu_2)/\sigma_j$, can be used in either experimental or nonexperimental designs and uses the standard deviation in one group as the standardizer. In an experimental design that includes a control group, the standard deviation in the control group is typically used as a standardizer. In an experimental design that compares a new treatment with a standard treatment, the standard deviation in the standard treatment group is typically used as a standardizer. In a nonexperimental design with unequal sample sizes, using the standard deviation from the largest group will give a narrower confidence interval. The `ci.stdmean2` function in the `statpsych` package will compute confidence intervals for all three types of standardized mean differences.

The estimates of δ and δ' have a slight positive bias in small samples. The bias can be reduced by multiplying $\hat{\delta}$ and $\hat{\delta}'$ by $1 - 3/[4(n_1 + n_2) - 9]$. The bias adjustment for δ'' is $1 - 3/[4n_j - 5]$ where n_j is the sample size of the group used to specify the standardizer. Researchers typically report a biased adjusted estimate of the standardized mean difference. However, the bias adjustment is not needed or recommended in the confidence interval computations.

Example 2.2. A random sample of 100 students diagnosed with test anxiety were randomly assigned to two treatment conditions (counseling method 1 and counseling method 2) and then completed a test anxiety assessment questionnaire following counseling. The sample means and standard deviations are given below ($n_1 = n_2 = 50$).

Treatment 1	Treatment 2
$\hat{\mu}_1 = 20.9$	$\hat{\mu}_2 = 19.1$
$\hat{\sigma}_1 = 3.85$	$\hat{\sigma}_2 = 3.19$

The 95% confidence interval for δ is [0.11, 0.91]. The researcher is 95% confident that in the study population of students with test anxiety, the mean test anxiety score would be 0.11 to 0.91 standard deviations greater if all students with test anxiety were given counseling method 1 rather than counseling method 2.

To interpret the confidence interval for δ in Example 2.2, imagine two normal curves – one curve for a population distribution of test anxiety scores for counseling method 1 and a second curve for a population distribution of test anxiety scores for counseling method 2. Now visualize the normal distribution for counseling method 1 shifted to the right of the normal distribution for counseling method 2 at least 0.11 standard deviations and at most 0.91 standard deviations. To visualize the magnitude of this shift, use the fact that one standard deviation from the mean is the point where the normal curve changes from concave down to concave up (the inflection point). Knowing that the point of inflection on a normal curve is one standard deviation from the mean, a researcher can then easily visualize two normal distributions that are separated by a specified number of standard deviations. Although the confidence intervals for δ and δ' do not require equal population variances, the *interpretations* of δ and δ' are most clear and useful when the variances are similar. The interpretation of δ'' is useful in applications where the variances are not similar.

In psychological research, values of standardized mean differences equal to about 0.2, 0.5, and 0.8 are typically described as "small", "medium", and "large" effect sizes, respectively. However, in some studies even a "small" effect size could have important scientific or practical implications. For example, an inexpensive mentoring program for first generation college students that would increase the population mean GPA by 0.2 standard deviations could also meaningfully increase their graduation rate and starting salaries.

A standardized mean difference can be transformed into a "common language effect size", denoted as π, where π is the proportion of people in the study population who would have a larger y score if they had received Treatment 1 rather than Treatment 2. In a nonexperimental design, π can be described as the probability of a randomly selected person from subpopulation 1 having a larger score than a randomly selected person from subpopulation 2.

A standardized mean difference can be transformed into π. To transform δ into π, simply find the area under the standard normal curve that is less than $\delta/\sqrt{2}$ (δ can be replaced with δ' or δ'' if equal population variances can be assumed). The `pnorm` function in R is useful for this purpose. This transformation can be applied to the lower confidence limit for δ and the upper confidence limit for δ to obtain a confidence interval for π. In Example 2.2, `pnorm`$(0.11/\sqrt{2})$ = .531 and `pnorm`$(0.91/\sqrt{2})$ = .740. We can be 95% confident that between 53.1% and 74.0% of all students in the study population would have a higher test anxiety score if they had all received counseling method 1 rather than counseling method 2.

2.5 Confidence Interval for a Ratio of Population Means

If the response variable is measured on a ratio scale, a ratio of population means μ_1/μ_2 is a unitless measure of effect size that could be more meaningful and easier to interpret than a standardized mean difference. An approximate $100(1-\alpha)\%$ confidence interval for μ_1/μ_2 that does not assume equal population variances is

$$exp\left[ln(\frac{\widehat{\mu}_1}{\widehat{\mu}_2}) \pm t_{\alpha/2;df}\sqrt{\frac{\widehat{\sigma}_1^2}{\widehat{\mu}_1^2 n_1} + \frac{\widehat{\sigma}_2^2}{\widehat{\mu}_2^2 n_2}}\right] \quad (2.6)$$

where $df = [\frac{\hat{\sigma}_1^2}{\hat{\mu}_1^2 n_1} + \frac{\hat{\sigma}_2^2}{\hat{\mu}_2^2 n_2}]^2 / [\frac{\hat{\sigma}_1^4}{\hat{\mu}_1^4 n_1^2 (n_1 - 1)} + \frac{\hat{\sigma}_2^4}{\hat{\mu}_2^4 n_2^2 (n_2 - 1)}]$. Ratios greater than 1 are usually easier to understand and describe than ratios less than 1. To convert a confidence for μ_1/μ_2 into a confidence interval for μ_2/μ_1 take the reciprocals of the confidence limits for μ_1/μ_2. The `ci.ratio.mean2` function in the `statpsych` package will compute Equation 2.6.

Suppose a 95% confidence interval for μ_1/μ_2 in a particular study is [1.51, 1.78]. This confidence interval has a simple interpretation: the researcher can be 95% confident that μ_1 is 1.51 to 1.78 times as large as μ_2.

2.6 Prediction Interval for a Score Difference

In some experiments, the researcher might want to estimate how the response variable score for *one* randomly selected person would differ under the two treatment conditions. For example, how much better will a particular child perform if given training method A instead of training method B? A $100(1-\alpha)\%$ prediction interval for this difference is

$$\hat{\mu}_1 - \hat{\mu}_2 \pm t_{\alpha/2;df} \sqrt{2\hat{\sigma}_p^2 + \hat{\sigma}_p^2/n_1 + \hat{\sigma}_p^2/n_2} \tag{2.7}$$

where $df = n_1 + n_2 - 2$ and $\hat{\sigma}_p^2$ is the pooled-variance estimate described previously. Equation 2.7 assumes equal population variances. There exists another version of this prediction interval that does not assume equal population variance but the *df* formula is complicated. A prediction interval for the difference in scores for one person will never be narrower, and is often much wider, than a confidence interval for the difference in population means. A prediction interval for a score difference, along with a confidence interval for $\mu_1 - \mu_2$, can provide a more realistic and useful assessment of a treatment effect. The `pi.score2` function in the `statpsych` package will compute Equation 2.7 and also the unequal variance version of Equation 2.7.

Example 2.3. In the two-group experiment described in *Example 2.1* where first-year college students were given either a cultural sensitivity reading list or a cultural sensitivity presentation with videos, the 95% confidence interval for $\mu_1 - \mu_2$ is [7.02, 9.18]. This confidence interval suggests that the *mean* number of culturally insensitive behaviors during a 2-month period would be 7.02 to 9.18 fewer if all students in the study population received the presentation with videos rather than the reading list. (*continued*)

Now the researcher wants to estimate how much better *one* randomly selected student from the study population would behave if given the presentation with videos rather than a reading list. The 95% prediction interval is [1.2, 14.9] and indicates that any one randomly selected student from the study population should exhibit about 1 to 15 fewer culturally insensitive behaviors during a 2-month period if given a cultural sensitivity video presentation rather than a cultural sensitivity reading list.

2.7 Directional Two-sided Test

In some studies, a theory will be supported if the results suggest that the population means are in the predicted direction (e.g., $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$). If a novel or unexpected directional effect is reported in the literature, a follow-up study might attempt to replicate the direction of the reported effect. As explained in Volume 5, directional tests can also be used to decide if it is appropriate to combine estimates of an effect that have been reported from multiple studies.

A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ can be used to perform a directional two-sided test of the following hypotheses.

H_0: $\mu_1 = \mu_2$ $\quad$ H_1: $\mu_1 > \mu_2$ $\quad$ H_2: $\mu_1 < \mu_2$

If the lower limit for $\mu_1 - \mu_2$ is greater than 0, reject H_0 and accept H_1: $\mu_1 > \mu_2$; if the upper limit for $\mu_1 - \mu_2$ is less than 0, reject H_0 and accept H_2: $\mu_1 < \mu_2$. The results are inconclusive if the confidence interval includes 0. Note that it is not necessary to develop special hypothesis testing rules for the value of δ (or δ' or δ'') because $\mu_1 = \mu_2$ implies $\delta = 0$, $\mu_1 > \mu_2$ implies $\delta > 0$, and $\mu_1 < \mu_2$ implies $\delta < 0$. The probability of making a directional error (accepting H_1 when H_2 is true or accepting H_2 when H_1 is true) is at most $\alpha/2$.

A directional two-sided test for the difference of two population means in a two-group design can also be performed using an *independent-samples t-test*. The following test statistic

$$t = (\hat{\mu}_1 - \hat{\mu}_2)/SE_{\hat{\mu}_1 - \hat{\mu}_2}$$

is used to select H_1 or H_2 using the following rules.

reject H_0 and accept H_1: $\mu_1 > \mu_2$ if $t > t_{\alpha/2;df}$

reject H_0 and accept H_2: $\mu_1 < \mu_2$ if $t < -t_{\alpha/2;df}$

fail to reject H_0 (i.e., an inconclusive result) if $|t| < t_{\alpha/2;df}$

The *t* statistic can be computed using either the separate-variance standard error or the pooled-variance standard error. SPSS and R will compute the *p*-value for the *t* statistic which obviates the need to determine the critical *t* value. If the *p*-value is less than α, then H_0 can be rejected and H_1 or H_2 can be accepted according to the sign of *t*. A *p*-value less than .05 does not imply that there is an important difference in the two population means, and a *p*-value greater than .05 does not imply that $\mu_1 = \mu_2$.

2.8 Equivalence Test

A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ can also be used to determine if μ_1 and μ_2 are similar. For example, a confidence interval for $\mu_1 - \mu_2$ can be used to choose between the following two hypotheses in an *equivalence test*

H_1: $|\mu_1 - \mu_2| < h$ $\qquad$ H_2: $|\mu_1 - \mu_2| \geq h$

where *h* is some value specified by the researcher. Usually, *h* represents a value of $\mu_1 - \mu_2$ that would be considered by experts to be small or unimportant. The interval *-h* to *h* is called the *region of practical equivalence*. If the confidence interval for $\mu_1 - \mu_2$ is completely contained within the range *–h* to *h*, then H_1 is accepted; if the confidence interval for $\mu_1 - \mu_2$ is completely outside the interval *-h* to *h*, then H_2 is accepted; otherwise, the results are inconclusive. The probability of falsely accepting H_1: $|\mu_1 - \mu_2| < h$ is at most $\alpha/2$. It is customary to use $\alpha = .10$ rather than $\alpha = .05$ in an equivalence test.

In applications where it is difficult to specify a value of $\mu_1 - \mu_2$ that would be considered by experts to be small or unimportant, it might be easier to specify a value of δ that would be considered small or unimportant. A confidence interval for δ can be used to choose between the following two hypotheses

H_1: $|\delta| < h$ $\qquad$ H_2: $|\delta| \geq h$

where h represents a value of δ that would be considered by experts to be small or unimportant. If the confidence interval for δ is completely contained within the interval $-h$ to h, then H_1 is accepted; if the confidence interval for δ is completely outside the interval $-h$ to h, then H_2 is accepted; otherwise, the results are inconclusive. Confidence intervals for δ' or δ'' can also be used to choose between H_1 and H_2.

Example 2.4. 180 Algebra II students were randomly divided into two groups of equal size. The first group received supplemental computer-assisted instruction, and the second group received supplemental instruction from a student teacher. At the end of the course, all 180 students were given the same final comprehensive examination which was scored from 0 to 100. If the difference in population means is between -4 and 4, the two instructional methods will be considered equivalent. The 95% confidence interval for $\mu_1 - \mu_2$ is [-3.7, 2.6] and suggests that the two supplemental teaching methods are equivalent in terms of their effectiveness.

2.9 Superiority and Noninferiority Tests

In a directional hypothesis test the goal is to accept either H_1: $\mu_1 - \mu_2 > 0$ or H_2: $\mu_1 - \mu_2 < 0$. If H_1 is accepted, then all that can be said is that μ_1 is greater than μ_2 although it is possible that μ_1 exceeds μ_2 by a trivial amount. In applications where a range of practical equivalence can be specified (and larger response variable scores are desirable), the following hypotheses are used to perform a *superiority test*.

H_0: $\mu_1 - \mu_2 = h$ $\quad$ H_1: $\mu_1 - \mu_2 > h$ $\quad$ H_2: $\mu_1 - \mu_2 < h$

If H_1: $\mu_1 - \mu_2 > h$ can be accepted, we can conclude that Treatment 1 is superior to Treatment 2 and the magnitude of the difference in population means is more than trivial. If H_2: $\mu_1 - \mu_2 < h$ is accepted, we conclude that Treatment 1 is not superior to Treatment 2.

A confidence interval for $\mu_1 - \mu_2$ can be used to implement a superiority test. If the lower limit for $\mu_1 - \mu_2$ is greater than h, reject H_0 and accept H_1: $\mu_1 - \mu_2 > h$ (i.e., Treatment 1 is superior to Treatment 2); if the upper limit for $\mu_1 - \mu_2$ is less than h, reject H_0 and accept H_2: $\mu_1 - \mu_2 < h$ (i.e., Treatment 1 is not superior to Treatment 2). The results are inconclusive if the confidence interval includes h.

In some applications the goal is to show that a less expensive treatment is not unacceptably worse than a more expensive treatment. Assuming larger response variable scores are desirable, the following hypotheses are used to perform a *noninferiority test*

$$H_0: \mu_1 - \mu_2 = -h \qquad H_1: \mu_1 - \mu_2 > -h \qquad H_2: \mu_1 - \mu_2 < -h$$

and a confidence interval for $\mu_1 - \mu_2$ can be used to implement this test. If the lower limit for $\mu_1 - \mu_2$ is greater than $-h$, reject H_0 and accept H_1: $\mu_1 - \mu_2 > -h$ (i.e., Treatment 1 is not inferior to Treatment 2); if the upper limit for $\mu_1 - \mu_2$ is less than $-h$, reject H_0 and accept H_2: $\mu_1 - \mu_2 < -h$ (i.e., Treatment 1 is inferior to Treatment 2). The results are inconclusive if the confidence interval includes $-h$.

A directional error in a noninferiority or superiority test occurs if H_1 is accepted when H_2 is true or if H_2 is accepted when H_1 is true. The probability of making a directional error in a noninferiority or superiority test is at most $\alpha/2$.

In applications where it is difficult to specify a value of $\mu_1 - \mu_2$ that would be considered small or unimportant, it might be easier to specify a region of practical equivalence for a superiority or noninferiority test in terms of a standardized mean difference.

***Example* 2.5.** A 90-minute cyber security online training program is currently required of all University of California faculty. A sample of 60 faculty were randomly selected from the UC faculty database and randomly divided into two groups of equal size. The first group was given a new and improved 25-minute online training program, and the second group was given the old 90-minute program. After training, a cyber security test (scored 0 to 50) was given to all 60 faculty. If the population mean test score for the 25-minute program is no more than 2 points worse than the population mean test score for the 90-minute program, then the 25-minute training program will be recommended for future use. The 95% confidence interval for $\mu_1 - \mu_2$ is [-0.7, 4.6] and suggests that the 25-minute training program is not inferior to the 90-minute training program and can be recommended for future use.

2.10 Variability Assessment

In addition to hypothesis tests or confidence intervals for population means or medians, it is informative to also examine confidence intervals for τ_j (Equation 1.14) and a confidence interval for τ_1/τ_2 where τ_j is the population MAD under

treatment *j* (in an experimental design) or subpopulation *j* (in a nonexperimental design).

In an experimental design, if one treatment causes an improvement in the population mean or median, the benefits of that treatment would be even more compelling if that treatment also caused a reduction in the variability of the response variable. Alternatively, if the variability is substantially greater for one of the treatment conditions, then that treatment might cause certain types of people to score higher and others to score lower. Future research could then focus on discovering which type of person is most affected or least affected by the treatment.

In nonexperimental designs, confidence intervals for τ_j and τ_1/τ_2 provide important information about individual differences within each subpopulation, and this information can help researchers avoid making a stereotyped description of some subpopulation of people based only on the subpopulation mean or median. For example, it might be the case that the *mean* performance on a certain task is greater for men than women, but it would be misleading to say "men perform better than women" on this task if there is substantial variability in the scores.

An approximate $100(1-\alpha)\%$ confidence interval for τ_1/τ_2 is

$$exp[ln(\frac{c_1\hat{\tau}_1}{c_2\hat{\tau}_2}) \pm z_{\alpha/2}SE_{ln(\hat{\tau}_1/\hat{\tau}_2)}] \tag{2.8}$$

where $SE_{ln(\hat{\tau}_1/\hat{\tau}_2)} = \sqrt{[(\hat{\mu}_1-\hat{\theta}_1)^2/\hat{\tau}_1^2 + \hat{\sigma}_1^2/\hat{\tau}_1^2 - 1]/n_1 + [(\hat{\mu}_2-\hat{\theta}_2)^2/\hat{\tau}_2^2 + \hat{\sigma}_2^2/\hat{\tau}_2^2 - 1]/n_2}$ and $c_j = n_j/(n_j-1)$. To convert a confidence for τ_1/τ_2 into a confidence interval for τ_2/τ_1 take the reciprocal of the confidence limits for τ_1/τ_2. The `ci.mad2` function in the `statpsych` package will compute Equation 2.8.

A confidence interval for τ_1/τ_2 can be used to decide if $\tau_1 > \tau_2$ or $\tau_1 < \tau_2$. If the lower limit is greater than 1, then accept $\tau_1 > \tau_2$; if the upper limit is less than 1 then accept $\tau_1 < \tau_2$; otherwise, the results are inconclusive. An inconclusive result does not justify the use of an equal-variance hypothesis test or confidence interval for $\mu_1 - \mu_2$. However, a narrow confidence interval for τ_1/τ_2 which suggests that

τ_1 is very similar to τ_2 could justify the use of the equal-variance hypothesis test or confidence interval for $\mu_1 - \mu_2$.

Unless the skewness or kurtosis of the response variable is substantially different in the two treatment conditions or the two subpopulations, τ_1/τ_2 will be approximately equal to σ_1/σ_2.

Example 2.6. Four hundred UCSC students were randomly selected from the psychology department's research participant pool and then randomly divided into two groups of equal size. The students were randomly paired to form 100 two-person teams within each group. Every team was given the same task, which involved searching the internet for specific facts about homelessness and then writing a brief proposal for assisting the homeless population in Santa Cruz. Each member of a two-person team worked in a separate room. The teams communicated using Zoom in group 1, and the teams communicated using email in group 2. After each team completed its task, the team proposal was graded on a 0 to 50 scale. The 95% confidence interval for $\mu_1 - \mu_2$ is [0.67, 2.52] indicating that the population mean score would be 0.67 to 2.52 points greater if the team members used Zoom rather than email. However, the 95% confidence interval for τ_1/τ_2 is [1.25, 2.65]. This result indicates that the population MAD of performance scores is 1.25 to 2.65 times as large in the Zoom condition than the email condition. The greater variability in performance scores using Zoom suggests that not all teams perform well using Zoom, and future research should attempt to identify the types of teams that perform best with Zoom or with email.

2.11 Assumptions

The separate-variance and pooled-variance confidence intervals and hypothesis tests for $\mu_1 - \mu_2$ assume random sampling and independence of observations. These confidence intervals also assume that the response variable has an approximate normal distribution in the study population within each treatment condition in an experimental design or within each subpopulation in a nonexperimental design.

The pooled-variance confidence interval and hypothesis test for $\mu_1 - \mu_2$ requires an additional assumption of equal population variances under each treatment condition (or within each subpopulation) and is called the *homoscedasticity assumption*. The pooled-variance confidence interval and test can be misleading, even with large sample sizes, when the population variances are unequal and the sample sizes are unequal. Specifically, the true coverage probability of a pooled-

variance $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ can be much less than $100(1-\alpha)\%$ when the larger sample size is used in the group with the smaller population variance. Alternatively, the true coverage probability of a pooled-variance $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ can be much greater than $100(1-\alpha)\%$ when the larger sample size is used in the group with the larger population variance. The pooled-variance hypothesis test and confidence interval for $\mu_1 - \mu_2$ can be used if the sample sizes are approximately equal; otherwise, the separate-variance hypothesis test and confidence interval for $\mu_1 - \mu_2$ is recommended. The `sim.ci.mean2` function in the `statpsych` package can be used to explore the performance of the pooled-variance and separate-variance confidence intervals for $\mu_1 - \mu_2$ with different sample sizes, different values of σ_1/σ_2, and different distribution shapes.

Violating the normality assumption is usually not a problem for hypothesis tests or confidence intervals regarding a difference or ratio of population means if the sample sizes are not too small ($n_j \geq 20$) and the response variable is at most moderately skewed. With approximately equal sample sizes and similar population variances, $100(1-\alpha)\%$ confidence intervals for $\mu_1 - \mu_2$ and μ_1/μ_2 will have true coverage probabilities close to $1-\alpha$ even when the response variable is highly skewed. However, if the sample sizes and variances are not similar, moderate skewness can adversely affect the confidence intervals for $\mu_1 - \mu_2$ and μ_1/μ_2 in small samples.

Confidence intervals for a standardized mean difference are sensitive to a violation of the normality assumption, and increasing the sample size will not mitigate the problem. However, the sensitivity to moderate nonnormality is minimal for small absolute values of δ, δ', or δ'' (less than .5) when the population variances are similar. Consequently, the normality assumption is not a concern when a confidence interval for δ, δ', or δ'' is used in an equivalence test where the standardized mean difference is assumed to be small. The confidence interval for δ'' is less sensitive to nonnormality than the confidence intervals for δ and δ'. If the standardized mean difference is not small, confidence intervals for δ and δ' should not be used unless the response variable or some transformation of the response variable is at most mildly nonnormal. The `sim.ci.stdmean2` function in the `statpsych` package can be used to explore the performance of the confidence

intervals for δ and δ'' with different sample sizes, different values of σ_1/σ_2, and different distribution shapes.

If the response variable is measured on a ratio scale, a confidence interval for μ_1/μ_2 is an attractive alternative to a confidence interval for a standardized mean difference because a confidence interval for μ_1/μ_2 is not sensitive to kurtosis or moderate skewness.

The prediction interval for a difference in scores (Equation 2.7) is sensitive to a violation of the normality assumption, and increasing the sample size will not mitigate the problem. Both skewness and kurtosis adversely affect the performance of the prediction interval. The prediction interval for a difference in scores should not be used unless the response variable or some transformation of the response variable is at most mildly nonnormal.

The confidence interval for τ_1/τ_2 assumes the response variable scores in each treatment condition or in each subpopulation have an approximate normal distribution. This assumption is not a concern if $n_j \geq 20$ and the response variable scores within each group are not highly nonnormal.

To informally assess the degree of nonnormality in a two-group design, subtract $\hat{\mu}_1$ from all of the group 1 scores and subtract $\hat{\mu}_2$ from all of the group 2 scores. Then estimate the skewness and kurtosis coefficients from these $n_1 + n_2$ deviation scores (these deviation scores are also called "residuals"). If the deviation scores are skewed, it may be possible to reduce the skewness by transforming (e.g., log, square-root, reciprocal) the response variable scores.

The values of μ_1 and μ_2 could be difficult to interpret if the response variable scores have been transformed in an effort to reduce nonnormality. Consequently, a confidence interval for $\mu_1 - \mu_2$ could be difficult to interpret and the researcher may want to report only a hypothesis testing result, which does not require an interpretation of the effect size magnitude. However, confidence intervals for δ, δ', and μ_1/μ_2 remain interpretable with transformed data because they are unitless measures of effect size.

2.12 Distribution-free Methods

If the response variable is highly skewed, a difference in population medians could be a more appropriate and meaningful measure of effect size than a difference in population means. An approximate $100(1-\alpha)\%$ confidence interval for $\theta_1 - \theta_2$ is

$$\hat{\theta}_1 - \hat{\theta}_2 \pm z_{\alpha/2}\sqrt{SE_{\hat{\theta}_1}^2 + SE_{\hat{\theta}_2}^2} \tag{2.9}$$

where $SE_{\hat{\theta}_j}^2$ is defined in Equation 1.10. This confidence interval assumes random sampling, independence among participants, and that tied scores are rare. Equation 2.9 can also be used to conduct a directional two-sided test, an equivalence test, a noninferiority test, or a superiority test. Equation 2.9 can be computed using the `ci.median2` function in the `statpsych` package.

The `sim.median2` function in the `statpsych` package can be used to examine the performance of the confidence interval for $\theta_1 - \theta_2$ with different sample sizes, different values of σ_1/σ_2, and different distribution shapes. It is instructive to compare the results with those from the `sim.mean2` function.

If the response variable is measured on a ratio scale, a ratio of population medians is a useful unitless measure of effect size. To obtain a confidence interval for θ_1/θ_2, compute Equation 2.9 from log-transformed response variable scores and then exponentiate the lower and upper limits. A confidence interval for θ_1/θ_2 can be computed using the `ci.ratio.median2` function in the `statpsych` package.

Equation 2.9 can be useful in time-to-event studies. If the study ends before some of the participants have exhibited the event, the scores for those participants will then be *right-censored*. Equation 2.9 requires $y_{(o_2)} \leq max$ in both groups where o_2 is defined in Equation 1.9 and *max* is the maximum possible time value for the study period. The `ci.ratio.median2` function can be used with right-censored data if the censored scores are set to some value greater than *max* and it has first been verified that $y_{(o_2)} \leq max$.

Example 2.7. A random sample of 40 opioid-addicted patients were random divided into two groups of equal size. The first group received cognitive behavior therapy (CBT) and the second group received medicated-assisted-treatment (MAT). The patients were followed for two years, and the number of days until relapse was determined for each patient. The relapse scores (in days) are given below.

CBT:	70	394	43	95	62	128	2	203	81	436	85	35	156	1	3	27	63	181	184	18
MAT:	102	120	78	78	417	124	86	171	176	129	230	194	100	157	306	411	164	103	193	312

The sample medians are 75.5 and 160.5 days for the CBT and MAT groups, respectively. A 95% confidence interval for $\theta_1 - \theta_2$ is [-159, -12]. The researcher can be 95% confident that the median time to relapse would be 12 to 159 days longer if all patients in the study population had been given MAT rather than CBT.

The *Mann-Whitney test* (also called the *Wilcoxon rank sum test*) is a distribution-free test of H_0: $\pi = .5$ where π is the common language effect size described in Section 2.4. In a two-group experiment, the Mann-Whitney test only assumes random sampling and independence among participants. SPSS and R will compute a *p*-value for the Mann-Whitney test that can be used to decide if the null hypothesis can be rejected. The Mann-Whitney test is usually a little less powerful than the independent-samples *t*-test, but it can be more powerful than the *t*-test if the response variable is highly leptokurtic. The Mann-Whitney test is usually more powerful than a directional two-sided test based on a confidence interval for a difference in medians using Equation 2.9.

In applications where the response variable does not have, and cannot be transformed to have, an approximate normal distribution, the standardized mean difference will be difficult to interpret because then it is not possible to visualize one standard deviation from the mean. In these situations, π is a useful measure of effect size because it has a clear and simple interpretation for any distribution shape. A distribution-free confidence interval for π does not have a simple formula but it can be computed using the `ci.mann` function in the `statpsych` package.

Example 2.8. A random sample of 20 male marijuana users was obtained from a university research participant pool and randomly divided into two groups of equal size. Group 1 was given a placebo, and group 2 was given 2.5 mg of edible THC. Amygdala activity levels of all participants were obtained while participants listened to an audio tape with high emotional content. The activity scores are shown below.

(*continued*)

THC:	9.4	10.3	58.3	106.0	31.0	46.2	12.0	19.0	135.0	159.0
Control:	14.6	5.1	8.1	22.7	6.4	4.4	19.0	3.2		

The activity scores for two participants in the control group were lost due to a random equipment malfunction. The sample medians are 7.25 and 38.6 for the THC and Control groups, respectively. A 95% confidence interval for $\theta_1 - \theta_2$ is [-22.4, 85.1]. This confidence interval includes 0 and H_0: $\theta_1 = \theta_2$ cannot be rejected. However, the 95% distribution-free confidence interval for π is [.629, 1.00]. This confidence interval excludes .5 and the Mann-Whitney null hypothesis (H_0: $\pi = .5$) can be rejected. The contradictory results are due to the fact that the Mann-Whitney test is usually more powerful than a test of H_0: $\theta_1 = \theta_2$ based on a confidence interval for $\theta_1 - \theta_2$. The researcher can be 95% confident that 62.9% to 100% all students in the research participant pool would have a higher activity score if they had all received THC rather than a placebo.

2.13 Sample Size Requirement for Desired Precision

The sample size requirement per group to estimate $\mu_1 - \mu_2$ with desired confidence and desired confidence interval width (w) is approximately

$$n_j = 8\tilde{\sigma}^2\left(\frac{z_{\alpha/2}}{w}\right)^2 + \frac{z_{\alpha/2}^2}{4} \tag{2.10}$$

where $\tilde{\sigma}^2$ is a planning value of the average within-group variance of the response variable for the two groups. The strategies for specifying a variance planning value described in section 1.29 can also be used to specify a within-group variance planning value. Equation 2.10 can be computed using the `size.ci.mean2` function in the `statpsych` package.

The sample size requirement per group to estimate δ or δ' with desired confidence and desired confidence interval width (w) is approximately

$$n_j = (\tilde{\delta}^2 + 8)\left(\frac{z_{\alpha/2}}{w}\right)^2 \tag{2.11}$$

where $\tilde{\delta}$ is a planning value of the standardized mean difference. The planning value can be specified using information from published research reports, a pilot study, or expert opinion. Equation 2.11 can be computed using the `size.ci.sdtmean2` function in the `statpsych` package. A larger value of $\tilde{\delta}^2$ will give a larger sample size requirement. Set $\tilde{\delta}^2$ to its largest likely value for a

conservatively large sample size requirement. Equation 2.11 can also be used to approximate the sample size required to estimate $\mu_1 - \mu_2$ if the variance planning value required in Equation 2.10 is difficult to specify.

The sample size requirement per group to estimate δ'' with desired confidence and desired confidence interval width is approximately

$$n_j = (2\tilde{\delta}''^2 + 8)(\frac{z_{\alpha/2}}{w})^2 \quad (2.12)$$

where $\tilde{\delta}''$ is a planning value of δ''. Unless $\tilde{\delta}''^2 = 0$, the sample size requirement for δ'' will be larger than the sample size requirement for δ or δ'. The `size.ci.sdtmean2` function will also compute Equation 2.12.

Specifying a narrow confidence interval width in Equations 2.10 - 2.12 might require a prohibitively large sample size. Rather than specifying a desirably narrow confidence interval width, the researcher could also try specifying the largest confidence interval width that would still provide useful information.

Example 2.9. A researcher wants to conduct a study to determine the effect of "achievement motivation" on the types of tasks one chooses to undertake. The study will ask participants to play a ring-toss game where they try to throw a small plastic ring over an upright post. The participants will choose how far away from the post they are when they make their tosses. The chosen distance from the post is the response variable. The independent variable is degree of achievement motivation (high or low) and will be manipulated by the type of instructions given to the participants. The results of a pilot study suggest that the standard deviation of the distance scores is about 9 inches within each condition. The researcher wants the 99% confidence interval for $\mu_1 - \mu_2$ to have a width of about 12 inches. The required sample size per group is approximately $n_j = 8(9^2)(2.58/12)^2 + 1.66 = 31.6 \approx 32$.

Example 2.10. A researcher will compare two methods of treating homelessness-induced PTSD in adolescents and will use a new measure of PTSD as the response variable. Given the novelty of the new PTSD measure, it is difficult for the researcher to specify a desired width of a confidence interval for $\mu_1 - \mu_2$. However, the researcher expects δ to be about 1.0 and would like a 95% confidence interval for δ to have a width of 0.5. The required sample size per group is approximately $n_j = (1^2 + 8)(1.96/0.5)^2 = 138.3 \approx 139$. The required sample size per group for δ'' is approximately $n_j = (2(1^2) + 8)(1.96/0.5)^2 = 153.6 \approx 154$.

It is often easier to specify the desired width of a confidence interval for a standardized mean difference than a confidence interval for $\mu_1 - \mu_2$. If the researcher plans to report a confidence interval for $\mu_1 - \mu_2$ and finds it easier to specify the desired confidence interval width for a standardized mean difference, then multiplying the desired confidence interval width for the standardized mean difference by $\sqrt{\tilde{\sigma}^2}$ gives the corresponding desired confidence interval width for $\mu_1 - \mu_2$ in Equation 2.10.

With a ratio-scale response variable, the sample size requirement per group to estimate μ_1/μ_2 with desired confidence and precision is approximately

$$n_j = 8\tilde{\sigma}^2 \left(\frac{1}{\tilde{\mu}_1^2} + \frac{1}{\tilde{\mu}_2^2}\right) \left[\frac{z_{\alpha/2}}{ln(\tilde{U}/\tilde{L})}\right]^2 + \frac{z_{\alpha/2}^2}{4} \qquad (2.13)$$

where $\tilde{\mu}_j$ is a planning value of μ_j, $\tilde{U}$ is the desired upper confidence limit, $\tilde{L}$ is the desired lower confidence limit, and $ln(\tilde{U}/\tilde{L})$ is the natural logarithm of $\tilde{U}/\tilde{L}$. For example, if μ_1/μ_2 is expected to about 1.3, the researcher may want the lower and upper confidence interval limits to be about 1.1 and 1.5 and $\tilde{U}/\tilde{L}$ would then be set to 1.5/1.1 = 1.36. Equation 2.13 can be computed using the `size.ci.ratio.mean2` function in the `statpsych` package. Note that Equation 2.13 requires planning values for μ_1 and μ_2 which limits its usefulness.

Example 2.11. A researcher will compare two methods of encouraging parents to read to their preschool children. The number of reading minutes per week is the response variable. The researcher plans to compute a 95% confidence interval for μ_1/μ_2 and would like the upper to lower interval endpoint ratio to be about 1.5. After reviewing the literature, the researcher set $\tilde{\sigma}^2$ = 200, $\tilde{\mu}_1^2$ = 50, and $\tilde{\mu}_2^2$ = 70. The required sample size per group is approximately $n_j = 8(200)(1/50^2 + 1/70^2)[1.96/ln(1.5)]^2 + 0.96 = 23.5 \approx 24$.

The `size.ci.median2` function in the `statpsych` package will compute the sample size required estimate $\theta_1 - \theta_2$ with desired confidence and desired confidence interval width. The required sample size depends on the shape of the distribution of y scores (see section 1.28).

2.14 Sample Size Requirement for Desired Power

The sample size requirement per group to perform a directional two-sided test for the difference in two population means with a specified α value and desired power is approximately

$$n_j = 2\tilde{\sigma}^2 \frac{\left(z_{\alpha/2} + z_\beta\right)^2}{(\tilde{\mu}_1 - \tilde{\mu}_2)^2} + \frac{z_{\alpha/2}^2}{4} \tag{2.14}$$

where $1 - \beta$ is the desired power of the test and $\tilde{\mu}_1 - \tilde{\mu}_2$ is a planning value of the anticipated effect size. Note that Equation 2.14 only requires a planning value for the difference in population means (i.e., the effect size) and does not require a planning value for each population mean.

In applications where it is difficult to specify $\tilde{\mu}_1 - \tilde{\mu}_2$ or $\tilde{\sigma}^2$, Equation 2.14 can be expressed in terms of a standardized mean difference planning value, as shown below.

$$n_j = \frac{2\left(z_{\alpha/2} + z_\beta\right)^2}{\tilde{\delta}^2} + \frac{z_{\alpha/2}^2}{4} \tag{2.15}$$

Equation 2.14 can be computed using the `size.test.mean2` function in the `statpsych` package. Equation 2.15 can also be computed using the `size.test.mean2` function by setting $\tilde{\mu}_1 = \tilde{\delta}$, $\tilde{\mu}_2 = 0$, and $\tilde{\sigma}^2 = 1$. SPSS has an option to compute the required sample size for desired power of an independent-samples *t*-test. The strategies for specifying an effect size described in section 1.29 can also be used to specify the effect size in Equations 2.14 and 2.15. In applications where the sample size is predetermined, the `power.mean2` function in the `statpsych` package can be used to estimate the power of the test for the given sample size, α value, and effect size.

Example 2.12. A researcher wants to compare two xenophobia treatments and wants the power of the test to be .9 with α = .05. The researcher expects the standardized mean difference to be about 0.5. The required number of participants per group is approximately $n_j = 2(1.96 + 1.28)^2/0.5^2 + 0.96 = 84.9 \approx 85$.

Example 2.13. A random sample of unhoused adults in Los Angeles will be randomly assigned to single-site or scattered-site permanent housing. A nine-item mental health questionnaire (scored 0 to 27) will be given to all participants after six months. The researcher sets $\tilde{\sigma}^2$ = 25 and expects a 3-point difference in the population mean mental health scores. For α = .05 and power of 1 – β = .95, the required number of unhoused individuals per group is approximately n_j = 2(25)(1.96 + 1.65)²/3² + 0.96 = 73.4 ≈ 74.

Let –*h* and *h* specify a region of practical equivalence. The sample size requirement per group to perform an equivalence test for the difference in two population means with a specified α value and desired power is approximately

$$n_j = 2\tilde{\sigma}^2 \frac{\left(z_\alpha + z_{\beta/2}\right)^2}{(h - |\tilde{\mu}_1 - \tilde{\mu}_2|)^2} + \frac{z_\alpha^2}{4} \tag{2.16}$$

where $|\tilde{\mu}_1 - \tilde{\mu}_2|$ is the expected effect size that must be smaller than *h*. Equivalence tests usually require very large sample sizes. Equation 2.16 can be computed using the `size.equiv.mean2` function in the `statpsych` package. In Equation 2.16, the power of the test is 1 – β which is defined as the probability of correctly accepting H1: $|\mu_1 - \mu_2| < h$.

Example 2.14. A researcher wants to show that women and men have similar means on a newly developed test of analytical reasoning. The test is scored on a 50 to 150 scale and the researcher believes that a 5-point difference in means would be small and unimportant. The required sample size per group to test H1: $|\mu_1 - \mu_2| < 5$ with power of .8, α = .10, an expected effect size of 1, and a standard deviation planning value of 15 is approximately n_j = 2(15²)(1.28 + 1.28)²/(1 – 5)² + 0.41 = 184.7 ≈ 185.

The sample size requirement per group to perform a superiority test for the difference in two population means with a specified α value and desired power is approximately

$$n_j = 2\tilde{\sigma}^2 \frac{\left(z_{\alpha/2} + z_\beta\right)^2}{(\tilde{\mu}_1 - \tilde{\mu}_2 - h)^2} + \frac{z_{\alpha/2}^2}{4} \tag{2.17}$$

where $\tilde{\mu}_1 - \tilde{\mu}_2$ is the expected effect size that must be greater than *h*. Equation 2.17 can be computed using the `size.supinf.mean2` function in the `statpsych` package. Equation 2.17 can also be used for a noninferiority test by replacing *h* with *-h* and specifying a value for $\tilde{\mu}_1 - \tilde{\mu}_2$ that is greater than *-h*.

Compared to the sample size requirement for a directional two-sided test where the goal, for example, is to show that $\mu_1 > \mu_2$, a larger sample is required to show that $\mu_1 - \mu_2 > h$ in a superiority test and a smaller sample size is required to show that $\mu_1 - \mu_2 > -h$ in a noninferiority test.

The sample size requirement per group to perform a Mann-Whitney with a specified α value and desired power is approximately

$$n_j = \frac{\left(z_{\alpha/2} + z_{\beta}\right)^2}{[6(\tilde{\pi} - .5)^2]} \tag{2.18}$$

where $\tilde{\pi}$ is a planning value of π. Recall that for experimental designs, π is the proportion of people in the study population who would have a larger y score if they had received Treatment 2 rather than Treatment 1. In a nonexperimental design, π can be thought of as the probability of a randomly selected person from the second subpopulation having a y score that is less than a randomly selected person from the first subpopulation. Equation 2.18 can be computed using the `size.test.mann` function in the `statpsych` package. The effect size for the Mann-Whitney test is $\tilde{\pi} - .5$ and setting this effect size to a minimally interesting value will give a conservatively large sample size requirement.

2.15 Unequal Sample Sizes

Using equal sample sizes has three major benefits: 1) if the population variances are approximately equal, confidence intervals are narrowest and hypothesis tests are most powerful when the sample sizes are equal, 2) when the pooled-variance confidence interval or hypothesis test is used, the effects of violating the equal variance assumption are minor when the sample sizes are equal, and 3) confidence intervals and hypothesis tests for a difference in population means are most robust to nonnormality when the sample sizes are equal. However, there are situations when equal sample sizes are less desirable. If one treatment is more expensive or risky than another treatment, the researcher might want to use fewer participants in the more expensive or risky treatment condition. Also, in experiments that include a control group, it could be easy and inexpensive to obtain a larger sample size for the control group.

The sample size formulas given above assume equal sample sizes per group. Suppose the researcher requires $n_2/n_1 = R$ (i.e., n_2 should be R times as large as n_1). The approximate sample size requirement for group 1 to estimate $\mu_1 - \mu_2$ with desired precision is

$$n_1 = 4\tilde{\sigma}^2(1 + 1/R)(\frac{z_{\alpha/2}}{w})^2 + \frac{z_{\alpha/2}^2}{4} \quad (2.19)$$

and the required sample size for group 1 to estimate δ with desired precision is

$$n_1 = 4[\tilde{\delta}^2(1 + 1/R)/8 + (1 + 1/R)](\frac{z_{\alpha/2}}{w})^2 \quad (2.20)$$

with n_2 set equal to n_1R.

To test H0: $\mu_1 = \mu_2$ with desired power, the approximate sample size requirement for group 1 is

$$n_1 = \tilde{\sigma}^2(1 + 1/R)(z_{\alpha/2} + z_\beta)^2/(\tilde{\mu}_1 - \tilde{\mu}_2)^2 + \frac{z_{\alpha/2}^2}{4} \quad (2.21)$$

or equivalently, in the case where $\tilde{\mu}_1 - \tilde{\mu}_2$ or $\tilde{\sigma}^2$ is difficult to specify,

$$n_1 = (1 + 1/R)(z_{\alpha/2} + z_\beta)^2/\tilde{\delta}^2 + \frac{z_{\alpha/2}^2}{4} \quad (2.22)$$

with n_2 set equal to n_1R. Equations 2.19 – 2.22 assume approximately equal population variances.

Example 2.15. A researcher wants to estimate $\mu_1 - \mu_2$ with 95% confidence and a desired confidence interval width of 2.5 with a variance planning value of 4.0. The researcher also wants n_2 to be 2 times the size of n_1. The sample size requirement for group 1 is approximately n_1 = 4(4.0)(1 + .5)(1.96/2.5)^2 + 0.96 = 15.7 ≈ 16 and then n_2 = 2(16) = 32 participants are required in group 2.

Example 2.16. A researcher wants to test H0: $\mu_1 = \mu_2$ with α = .05 and power of .95. The researcher also wants n_2 to be about one-fourth the size of n_1. The researcher expects the standardized mean difference to be about 0.75. The sample size requirement for group 1 is approximately n_1 = (1 + 1/0.25)(1.96 + 1.65)^2 /0.75^2 + 0.96 = 165.8 ≈ 117 and then n_2 = (1/4)(117) ≈ 30 participants are required in group 2.

2.16 Data Visualization

The estimated means or medians for each group can be presented graphically using a *bar chart.* A bar chart for two groups consists of two bars, one for each group, with the height of each bar representing the value of the sample mean or median. Bar charts of parameter estimates can be misleading because the estimates contain sampling error of unknown magnitude and direction. There is a tendency to incorrectly interpret a difference in bar heights as representing a difference in population value. This misinterpretation can be avoided by graphically presenting the imprecision of the parameter estimates with 95% confidence interval lines as shown in the bar chart below.

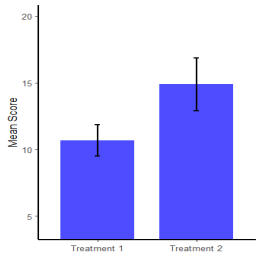


Figure 2.1 Bar chart for a two-group experiment

An *error bar* chart is an alternative to a bar chart with confidence interval lines. An error bar chart displays the estimated means or medians with a 95% confidence interval line as shown below. Error bar charts are sometimes easier to read than bar charts.

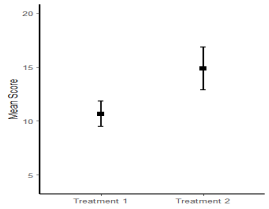


Figure 2.2 Error bar chart for a two-group experiment

Bar charts with confidence interval lines and error bar charts are sometimes used for visual hypothesis testing, but researchers frequently do this incorrectly. Non-overlapping 95% confidence interval lines for the individual means implies a

p-value less than .05 for a directional two-sided test. However, overlapping confidence interval lines do not necessarily imply a *p*-value greater than .05 for a directional two-sided test. With approximately equal sample sizes and standard deviations, if the amount of overlap is less than about one-half the length of the 95% confidence interval lines, then the *p*-value can still be less than .05.

The confidence interval lines provide useful information in both experimental and nonexperimental designs. In an experimental design, the confidence interval line for a treatment *j* describes a range of plausible values for the population mean assuming every person in the population had received treatment *j*. In a nonexperimental design, the confidence interval line for group *j* describes a range of plausible values for the population mean in subpopulation *j*.

The scale of the vertical axis should be selected in a way that does not give an exaggerated impression of the mean or median difference. Suppose the response variable is measured on a 1 to 50 scale and the sample means are $\hat{\mu}_1$ = 25.4 and $\hat{\mu}_1$ = 25.6. This small difference in sample means will appear misleadingly large in a bar chart where the minimum and maximum values of vertical axis have been set to 25 and 26. If the response variable is measure on a 1 to 5 or 1 to 7 Likert scale, setting the vertical axis range equal to the response variable range can avoid a misleading impression of the mean difference. In other situations, the minimum value of the vertical axis could be set to the 5th percentile of the sample y scores and the maximum value of the vertical axis could be set to the 95th percentile of the sample y scores. Using the 5th and 95th percentiles for the range of the y axis has the added benefit of providing additional information about the variability of the y scores in the study.

SPSS has an option to produce bar charts with "Standard Error" lines that are shorter than 95% confidence interval line. Standard error lines are approximate 68% confidence interval lines and their use is not recommended.

A dot plot and a box plot of the y scores within each group are useful ways to visualize the data in the data screening process. A *side-by-side box plot* is a plot of box plots for two or more groups. An example of a side-by-side box plot for a two-group experiment is shown below. This plot indicates that there is an outlier in Treatment 1 that requires investigation.

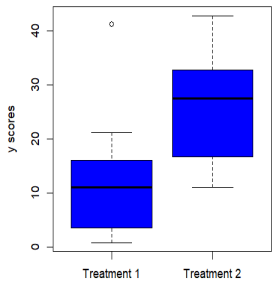


Figure 2.3 Side-by-side box plot for a two-group experiment

2.17 Missing Data

In some experiments, certain types of participants are more likely to withdraw from one treatment than another treatment. In the context of an experimental design, this type of missing data will be referred to here as *differential nonrandom attrition*. With differential nonrandom attrition, participants who complete the study could differ across treatment conditions in terms of some attribute that has a causal effect on the response variable. That attribute would then be a confounding variable. In these situations, we say that the *internal validity* of the experiment has been compromised because we can no longer assume a causal relation between the independent variable and the dependent variable. Furthermore, the remaining sample of participants is no longer a random sample from the study population. If one or more confounding variables can be identified, the statistical methods in Volume 2 can be used to improve the interval validity of the study.

Now consider an experiment where a certain type of participant is more likely to drop out of the study but the dropout rate is about the same in each treatment condition. In this situation, we might be able to argue that the loss of participants is not related to the independent variable and the internal validity of the experiment should not be compromised. However, the *external validity* of the experiment will be compromised. External validity is the extent to which the results of a study can be generalized to different types of participants. In the context of an experimental design, this type of missing data is referred to here as *non-differential nonrandom attrition*. With non-differential nonrandom attrition, the participants who complete the study are no longer a random sample from the

original study population. The remaining participants could be assumed to be a random sample from a smaller study population of participants who would have completed the study. This change in the size and nature of the study population decreases the external validity of the study. Furthermore, a loss of data will reduce the power of the hypothesis tests and will increase the widths of the confidence intervals.

In some experiments, the data for a few participants might be lost due to equipment failure, a lost questionnaire, or a participant withdrawing from the study due to illness. In these cases, where the missing data are assumed to be the result of purely random influences, we say that the missing data are *missing completely at random* (MCAR). If the missing data are MCAR, neither the internal validity nor the external validity of the study will be compromised. Furthermore, the remaining sample will still be a random sample from the study population. However, the loss of data will reduce the power of a hypothesis test and will increase the width of a confidence interval. If possible, the researcher might be able to obtain a second small random sample from the same study population and randomly assign the new participants to the levels of the independent variable. The new random sample would be combined with the original random sample and then the analyses would be performed on the combined sample.

A sample size that is computed from a sample size formula can be adjusted upwards to account for the possibility of missing data to help achieve the desired precision or power. However, increasing the sample size will not correct the threats to internal validity due to differential nonrandom attrition or improve the reduction in external validity to due to non-differential nonrandom attrition.

2.18 Multiple Response Variables

All of the methods described in this chapter and the other three chapters are illustrated using a single response variable. In some studies, the analysis of two or more response variables may be required to answer certain research questions. For example, in a one-group design a researcher may want to estimate the mean scores of five personality traits (e.g., neuroticism, extroversion, openness, agreeableness, conscientiousness) in a study population of gambling addicts. Or in a two-group design, a researcher may want to determine if the means of resilience scores and

ambition scores are both greater in a study population of first-generation college students than in a study population of continuing-generation college students.

For any single $100(1-\alpha)\%$ confidence interval, we can be $100(1-\alpha)\%$ confident that the confidence interval has included the population parameter value (if all assumptions have been satisfied). However, if a $100(1-\alpha)\%$ confidence interval is computed for r different response variables, it can be shown that we can be *at least* $100(1-r\alpha)\%$ confident that *all* r confidence intervals have included their population parameters. For example, if five 95% confidence intervals are computed, we can be at least $100(1-r\alpha)\%$ = 100(1 – .25)% = 75% confident that all five confidence intervals have included their population parameter values.

Researchers usually want to be at least $100(1-\alpha)\%$ confident, rather than at least $100(1-r\alpha)\%$ confident, that all r confidence intervals have included their population parameters. One simple way to achieve this objective is to use $\alpha^* = \alpha/r$ rather than α in the critical t-value (in Equation 2.1) or critical z-value (in Equations 2.2 or 2.6) for each confidence interval. This type of adjustment to the alpha value is called a *Bonferroni* adjustment. Multiple confidence intervals that use a Bonferroni adjustment are called *simultaneous confidence intervals*.

Example 2.17. Eighty randomly selected middle school children were randomly divided into two groups of equal size. All 80 participants watched one of two different 5-minute anti-smoking videos that described the negative health consequences of smoking. Group 1 viewed video 1 and group 2 viewed video 2. After viewing a video, all participants were given a quiz to assess their knowledge of the negative health consequences of smoking. The participants were also given an "attitude towards smokers" questionnaire (higher scores represent greater negative attitudes) and an "intention to smoke" questionnaire (lower scores represent less intention). Simultaneous 95% confidence intervals were computed for the differences in population means (video 1 vs. video 2) for each of three response variables. Using α^* = .05/3 = .0167, the simultaneous 95% confidence interval results for knowledge, attitude, and intention are [5.1, 9.2], [2.5, 7.6], and [-4.9, -1.3], respectively. These results indicate that video 1 is more effective than video 2 in communicating the negative health consequences of smoking, producing more negative attitudes towards smokers, and reducing the intention to smoke.

2.19 Exploratory Data Analysis

In a *confirmatory* study, every aspect of the data analysis has been decided prior to the collection of the data. In an *exploratory* study, the researcher might examine several different response variables without a Bonferroni correction, try different transformations of a response variable, and run more than one type of hypothesis test in the hopes of finding a "significant" result. A $100(1 - \alpha)\%$ confidence interval will include the population parameter value with probability $1 - \alpha$ if all assumptions have been satisfied and the study is confirmatory. Likewise, a two-sided directional hypothesis test will lead to an incorrect decision with probability that is at most $\alpha/2$ if all assumptions have been satisfied and the study is confirmatory. In an exploratory study, the probability of a $100(1 - \alpha)\%$ confidence interval including the population parameter can be less than $1 - \alpha$, and the probability of making a directional error can be greater than $\alpha/2$.

Describing the results of an exploratory study as if it was a confirmatory study is a leading cause of the "replication crisis" in psychology. An increasing number of psychology journals are recommending a public *preregistration* of research plans prior to data analysis. In a preregistration, the researcher specifies the planned sample size, the planned data collection procedure, the planned research design and materials, and all planned statistical analyses. This information is then posted in a time-stamped online repository that can be accessed by editors, reviewers, and other researchers.

2.20 Ethical Issues

Any study that uses human subjects should advance knowledge and potentially lead to improvements in the quality of peoples' lives – but the researcher also has an obligation to protect the rights and welfare of the participants in the study. These two goals are often in conflict and lead to ethical dilemmas. The most widely used approach to resolving ethical dilemmas is to weigh the potential benefits of the research against the costs to the participants. Evaluating the costs and benefits of a proposed research project that involves human subjects can be extremely difficult, and this task is assigned to the *Institutional Review Board* (IRB) at most universities. Researchers who plan to use human subjects in their research must

submit a written proposal to the IRB for approval. The IRB will carefully examine research proposals in terms of the following issues:

Informed consent – Will participants be informed of the nature of the study, will they explicitly agree to participate, and can they decline to participate?

Coercion to participate – Will participants be coerced into participating or be offered excessive inducements?

Confidentiality – Will the data collected from participants be used only for research purposes and not divulged to others?

Physical and mental stress – Does the study involve more than *minimal risk*? Minimal risk is defined as a risk that is no greater in probability or severity than ordinarily encountered in daily life or during a routine physical or psychological exam.

Deception – Is deception needed in the study? If deception is used, are participants debriefed after the study? *Debriefing* is used to clarify the nature of the study to the participants and reduce any stress or anxiety experience by the participants that was could have been caused by the deception.

In addition to the principles governing the treatment of human subjects, researchers are bound by a set of ethical standards. Violation of these standards is called *scientific misconduct*. There are three basic types of scientific misconduct:

Scientific dishonesty – Examples include: the fabrication or falsification of data and *plagiarism.* Plagiarism is the use of another person's ideas, processes, results, or words without giving appropriate credit.

Unethical behavior – Examples include: sexual harassment of research assistants or research participants; abuse of authority; failure to follow university or government regulations; and inappropriately including or excluding authors on a research report or conference presentation.

Deceptive research practices – Examples include: performing an exploratory analysis and describing the results as confirmatory; failure to assess critical assumptions for statistical tests or confidence intervals; claiming to have predicted an unexpected finding; deleting legitimate data in an effort to obtain desired results; presenting an inconclusive result as if it supports the researcher's hypothesis of a zero effect; using language in the written report that downplays negative results; and exaggerating the generality of the results.

In experimental designs there are ethical costs associated with randomly assigning participants to treatment and control groups where the treatment is expected to be beneficial. The goal of these studies is to estimate the magnitude of the beneficial effect rather than test a directional two-sided hypothesis. One way to reduce the ethical costs in this type of experiment is to use a *waitlist control group* where the control group is compared with the treated group. Then, after all the analyses have been completed, the control group also receives the beneficial treatment.

In experiments where two treatments will be assessed and one treatment is believed to be more beneficial than the other treatment, ethical costs can be reduced by randomly assigning more participants to the more beneficial treatment and fewer participants to the less beneficial treatment.

A *Zelen design* can be used to reduce the ethical costs of an experiment if the only goal of the study is to determine the direction of the treatment effect (using a directional two-sided test) and not to estimate the magnitude of the treatment effect. In a Zelen design, participants are randomly assigned to a standard treatment or a new treatment, but participants in the new treatment group are given the option to receive the standard treatment. A directional two-sided test compares the two original groups formed by random assignment even though some of the participants in new experimental treatment group will have opted to receive the standard treatment. This contamination of treatments will reduce the magnitude of $\mu_1 - \mu_2$ (and hence the power of the test) but will not change the sign of $\mu_1 - \mu_2$. The Zelen design requires a larger sample size to compensate for the reduction in effect size due to treatment contamination.

Another strategy for reducing the ethical costs of an experiment is to ask participants if they are willing or unwilling to be randomly assigned to treatment conditions. The unwilling participants can choose which treatment condition they prefer and the willing participants are randomly assigned to treatment conditions. The resulting design is a two-factor design with "self-select" and "randomize" as the two levels of one factor and the treatment conditions as levels of the second factor. Two-factor designs are described in Chapter 3.

The ethical costs associated with randomly assigning participants to a control group or a potentially beneficial treatment group can be eliminated using the pretest-posttest design described in Chapter 4 where all participants are measured on one or more occasions prior to treatment and then on one or more occasions following treatment.

Key Terms

independent variable
dependent variable
experimental design
control group
effect size
nonexperimental design
stratified random sampling
confounding variable
subpopulation
error variance
Cohen's *d*
independent-samples *t*-test
equivalence test
superiority test
noninferiority test
homoscedasticity assumption
Mann-Whitney test
bar chart
internal validity
differential nonrandom attrition
external validity
non-differential nonrandom attrition
random attrition
Bonferroni adjustment
simultaneous confidence intervals
informed consent
minimal risk
debriefing
scientific misconduct
deceptive research practices
waitlist control group
exploratory data analysis
preregistration
side-by-side box plot

Concept Questions

1. What are the three conditions that must be satisfied to show that one variable is causally related to another variable?

2. Explain how the results of an experiment can provide evidence that the independent variable has a causal effect on the response variable.

3. How can a confidence interval for $\mu_1 - \mu_2$ be used to test H_0: $\mu_1 - \mu_2 = 0$ and then select H_1: $\mu_1 - \mu_2 > 0$ or H_2: $\mu_1 - \mu_2 < 0$?

4. How can a confidence interval for δ be used to test H_1: $|\delta| < 0.1$ against H_2: $|\delta| \geq 0.1$?

5. How can a confidence interval for $\mu_1 - \mu_2$ be used to select H_1: $|\mu_1 - \mu_2| < 2.5$ or H_2: $|\mu_1 - \mu_2| \geq 2.5$?

6. How can a confidence interval for $\mu_1 - \mu_2$ be used to test H_0: $\mu_1 - \mu_2 = -5$ and then select H_1: $\mu_1 - \mu_2 > -5$ or H_2: $\mu_1 - \mu_2 < -5$?

7. What information can be obtained from a confidence interval for $\mu_1 - \mu_2$ that cannot be obtained from an independent-samples *t*-test?

8. When would a researcher prefer to report a confidence interval for a standardized mean difference rather than an unstandardized mean difference?

9. What does it mean when a researcher declares the results of an independent-samples *t*-test to be "significant"? How should a "nonsignificant" result be interpreted?

10. What are the assumptions for a confidence interval or test for a difference in population means using the pooled-variance method?

11. What are the assumptions for a confidence interval or test for a difference in population means using the unequal-variance method?

12. Explain how μ_1 and μ_2 are interpreted differently in an experimental design compared to a nonexperimental design.

13. What are the consequences of participants randomly dropping out of both groups in a two-group experiment?

14. What are the consequences of certain types of participants dropping out only from one group in a two-group experiment?

15. What are the consequences of certain types of participants dropping out equally from both groups in a two-group experiment?

16. What are the advantages and disadvantages of sampling from a small study population?

17. Why is a violation of the normality assumption not a major concern when testing or estimating $\mu_1 - \mu_2$ when the sample sizes are not small?

18. Why are confidence interval lines recommended in bar charts?

19. How does the value of $\tilde{\mu}_1 - \tilde{\mu}_2$ affect the sample size requirement for testing H_0: $\mu_1 = \mu_2$ with desired power?

20. When would a Mann-Whitney test be preferred to an independent-samples *t*-test?

21. When would a confidence interval for a difference in population medians be preferred to a confidence interval for a difference in population means?

22. What useful information could be obtained by examining a confidence interval for a ratio of population MADs in a two-group design?

23. A random sample of 30 students was obtained from a research participant pool of about 5,000 undergraduates at UC Santa Barbara. The sample was randomly divided into two groups of equal size. Group 1 received a daily meditation app, and Group 2 received a daily planner app. After two weeks, the Perceived Stress Scale (scored from 0 to 40 with higher scores representing greater stress) was given to all 30 participants. A 95% confidence interval for $\mu_1 - \mu_2$ was [-3.4, -0.7]. Interpret this result.

24. Explain why each of the following interpretations of [-3.4, -0.7] in the above example is incorrect or inadequate.

a) We are 95% confident that the mean stress score of the 30 students would be between 0.7 and 3.4 greater if they had used the daily planner app rather than the daily meditation app.

b) We are 95% confident that the mediation app will reduce a student's stress score between 0.7 and 3.4 points.

c) We are 95% confident that the difference in population means for the two conditions would be between .07 and 3.4.

d) We are 95% confident that $\mu_1 - \mu_2$ is between -3.4 and -0.7.

Data Analysis Problems

2-1. There are more than 8,000 teenage driver fatalities in the U.S. each year. A new driver training simulator has been developed to provide training in basic driving skills, night driving, and hazard detection. If the training simulator is shown to be effective, researchers will begin work on a home version of the simulator. A random sample of 20 high school students was obtained from a list of about 6,000 Fresno County students enrolled in a driver's education course. The students were randomly assigned to two groups of equal size. One group received the traditional driver's education course, and the second group was given 35 hours of simulator training. After training, all 20 students were given a driving skills exam that was scored on a 0 to 50 point scale. The driving skills scores are shown below.

Group 1:	32	39	26	35	43	27	40	37	34	29
Group 2:	36	44	47	42	49	39	46	31	33	48

a) Describe the study population.

b) Describe the population means (μ_1 and μ_2) in the context of this study.

c) Use SPSS or R to test H_0: $\mu_1 = \mu_2$ with $\alpha = .05$. Report *t*, *df*, and *p*-value in APA style and state your conclusion. Do not assume equal population variances.

d) Use SPSS or R to compute a 95% confidence interval for $\mu_1 - \mu_2$ and interpret the result. Do not assume equal population variances.

e) The additional simulation training will be considered superior to the traditional driver's education course if $\mu_2 - \mu_1 > 1$. Use the 95% confidence interval for $\mu_1 - \mu_2$ to decide if the additional simulator training is superior to the traditional course.

f) Use the `pi.score2` function in the `statpsych` package to compute a 95% prediction interval for the difference in scores under the two training methods for a single student. Do not assume equal variances. Interpret the result.

g) Use SPSS or R to produce a bar chart for the two means with 95% confidence interval lines for the population means.

h) The researcher wants to conduct a similar study in a larger city such as Chicago. How many driver education students per group (with equal sample sizes) would be needed to perform a directional two-sided test with power of .9 at α = .05 assuming a 5-point difference in population means? Use the average of the sample variances from this experiment as your planning value of the average within-group error variance.

2-2. A widely used scale to measure a person's "character" contains a "courage" subscale. Women tend to score lower on this subscale, which is inconsistent with psychological theory. A new courage subscale was developed using only gender-neutral terms. A stratified random sample of 200 men and 200 women was obtained from a labor union directory of about 160,000 men and 140,000 women. The 400 participants were asked to answer the new 10-item courage subscale which is scored on a 1 to 50 scale. The sample means and standard deviations of the courage subscale scores are given below.

Men	Women
$\hat{\mu}_1$ = **30.2**	$\hat{\mu}_2$ = **30.8**
$\hat{\sigma}_1$ = **10.5**	$\hat{\sigma}_2$ = **11.2**

a) Describe the two study populations.

b) Describe the population means (μ_1 and μ_2) in the context of this study.

c) Use the `ci.mean2` function in the `statpsych` package to compute a 95% confidence interval for $\mu_1 - \mu_2$ and interpret the result. Do not assume equal population variances.

d) Use the `ci.stdmean.strat` function in the `statpsych` package to compute a 95% confidence interval for the population standardized mean difference and interpret the result (set p1 = 160,000/300,000 = .533).

e) The researchers will claim that the new courage subscale is not gender biased if they can accept H_1: $|\mu_1 - \mu_2| < 3$ with α = .05. Use the 95% confidence interval for $\mu_1 - \mu_2$ to decide if H_1: $|\mu_1 - \mu_2| < 3$ can be accepted and state your conclusion.

2-3. Crime suspects often confess under lengthy interrogation and then later recant their confession. Some juries are more willing to ignore the original confession, but others are not. A researcher suspects that the physical characteristics of the defendant might affect a juror's willingness to ignore a recanted confession. Thirty college students were randomly selected from a university research participant pool of about 1,000 undergraduate students. The 30 participants were randomly divided into two groups of equal size. Each participant read one of two vignettes about a hit-and-run driver who was harshly interrogated by police for five hours and then confessed to the crime. The following day, the suspect recanted his confession. In both vignettes the driver was described as a white 22-year-old male with brown eyes and brown hair. The first group read vignette 1 where the driver was described as "average height and frail". The second group read vignette 2 where the driver was described as "tall and muscular". (*continued*)
The participants were asked to imagine themselves as jurors in this hit-and-run trail and to rate on a 1 to 10 scale their willingness to ignore the original confession (where 1 represent extremely unwilling and 10 represents extremely willing). Their scores are shown below.

Group 1:	6	9	9	7	8	8	10	9	5	7	5	6	7	8	5
Group 2:	3	9	2	5	4	7	4	3	3	2	4	1	4	5	2

a) Describe the study population.

b) Describe the population means (μ_1 and μ_2) in the context of this study.

c) Use SPSS or R to compute a 95% confidence interval for $\mu_1 - \mu_2$ and interpret the result. Do not assume equal population variances.

d) Use SPSS or R to perform a directional two-sided test with α = .05. Report *t*, *df*, and *p*-value in APA style and state your conclusion. Do not assume equal population variances.

e) Use the `ci.stdmean2` function in the `statpsych` package compute a 95% confidence interval for the population standardized mean difference (unweighted variance standardizer) and interpret the result.

f) Use SPSS or R to perform a Mann-Whitney test and report the *p*-value.

g) The researcher wants to replicate this study using a larger sample size at a different university. How many participants are needed per group (with equal sample sizes) to obtain a 95% confidence interval for $\mu_1 - \mu_2$ that has a width of 2? Use the average of the sample variances from this experiment as your planning value for the average within-group error variance.

Suggested Readings

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge. **Chapters 6 and 11**

Howell, D. C. (2013). *Statistical methods for psychology (8th ed.)* Thompson & Wadsworth. **Sections 7.5, 7.6, and 7.7**

Hollander, M. & Wolfe, D. A. (1999). *Nonparametric statistical methods (2nd ed.).* Wiley. **Chapter 4**

Snedecor, G. W. & Cochran, W. G. (1989). *Statistical methods (8th ed.).* Iowa State Press. **Sections 6.5 – 6.12, and Section 6.14**

Chapter 3

Between-subjects Designs

3.1 One-factor Experimental Designs

In a one-factor experimental design, the randomly selected participants are randomly assigned to $a \geq 2$ groups. Each group receives one of the levels of the independent variable, with participants being treated identically in every other respect. It is common, but not necessary, to have an equal number of participants in each group. The two-group experimental design described in Chapter 2 is a special case of a one-factor experimental design with $a = 2$. The groups represent the *levels* of the independent variable. An independent variable with a specified number of levels is referred to as a "factor", and when participants are randomly assigned to the levels of a factor, the factor is referred to as a "treatment factor". In Chapter 2 and Chapter 3, the factors are referred to as "between-subjects" factors because each factor level consists of a different group of participants. Designs with "within-subjects" factors, where each participant receives all levels of the factor, are introduced in Chapter 4.

The population parameters that can be estimated in a one-factor experiment are $\mu_1, \mu_2, \ldots, \mu_a$ where μ_j ($j = 1$ to a) is the population mean of the response variable if all members of the study population had received level j of the independent variable. One method of assessing the differences among the population means is to compute confidence intervals for all possible pairs of differences. For example, with $a = 3$ the following *pairwise comparisons* of population means could be examined.

$$\mu_1 - \mu_2 \qquad \mu_1 - \mu_3 \qquad \mu_2 - \mu_3$$

In a one-factor experiment with a levels there are $a(a-1)/2$ pairwise comparisons. Confidence intervals for any of the two-group measures of effect size (e.g., mean difference, standardized mean difference, mean ratio, median difference, median ratio) described in Chapter 2 can be used to analyze any pair of groups in a one-factor experimental design.

Simultaneous confidence intervals for g pairwise comparisons can be obtained by using a Bonferroni adjustment $\alpha^* = \alpha/g$ rather than α in the critical t-value or critical z-value for each confidence interval. In the special case where the researcher will examine *all* $a(a-1)/2$ pairwise differences in population means, the *Tukey-Kramer* method can be used to obtain simultaneous confidence intervals that are narrower than the Bonferroni confidence intervals.

The classical Tukey-Kramer method for comparing all possible pairs of means assumes equal population variances, but a version of the Tukey-Kramer method that does not require equal population variances can be computed using the `ci.tukey` function in the `statpsych` package. SPSS provides an option to compute *Games-Howell* confidence intervals for all pairwise comparisons of means that are the same as the unequal variance version of the Tukey-Kramer confidence intervals. The Tukey-Kramer and Games-Howell methods are used only when the researcher is interested in examining all possible pairwise differences.

A Bonferroni confidence interval will be narrower than a Tukey-Kramer or Games-Howell confidence interval if, *prior* to an examination of the sample results, the researcher is interested in only a *subset* of the $a(a-1)/2$ pairwise comparisons. These types of comparisons are called *planned comparisons*. For g planned comparisons, the Bonferroni adjustment is $\alpha^* = \alpha/g$. However, if g of the $a(a-1)/2$ possible pairwise comparisons appear interesting *after* an examination of the sample results, Bonferroni confidence intervals with $\alpha^* = \alpha/g$ are not appropriate and it is necessary to use Tukey-Kramer or Games-Howell confidence intervals.

Example 3.1. One hundred and twenty students were randomly selected from a list of about 5,400 undergraduate students. The 120 students were randomly divided into three groups of equal size, and all 120 students were given a nonverbal intelligence test (Raven's Progressive Matrices) under identical testing conditions. The raw scores for this test range from 0 to 60. The students in group 1 were told that they were taking a "very difficult intelligence test". The students in group 2 were told to solve a set of "interesting puzzles". The students in group 3 were not told anything. Simultaneous Tukey-Kramer confidence intervals for all pairwise comparisons of population means are given below.

Comparison	95% Lower Limit	95% Upper Limit
$\mu_1 - \mu_2$	-5.4	-3.1
$\mu_1 - \mu_3$	-3.2	-1.4
$\mu_2 - \mu_3$	1.2	3.5

(*continued*)

The researcher is 95% confident that the mean intelligence score would be 3.1 to 5.4 greater if all 5,400 undergraduates had been told to solve puzzles instead of taking a difficult IQ test, 1.4 to 3.2 greater if they all had been told nothing instead of being told they were taking a difficult IQ test, and 1.2 to 3.5 greater if they all had been told to solve puzzles instead of being told nothing. The simultaneous confidence intervals allow the researcher to be 95% confident regarding all three of these conclusions.

3.2 Classification Factors

Instead of randomly assigning participants to the levels of a treatment factor, a random sample of participants could be classified into $a \geq 2$ groups according to some pre-existing characteristic (e.g., ethnicity, gender, political affiliation). These groups define the levels of a *between-subjects classification factor*. As explained in Section 2.2, either simple random sampling or stratified random sampling can be used to obtain a sample of participants within each level of the classification factor. If simple random sampling is used, the number of participants in each group will not be known in advance and the sample sizes in some groups could be very small. Stratified random sampling is usually more difficult to implement, but it has the advantage of obtaining the desired number of participants in each group. A study with a single classification factor is a nonexperimental design and is a generalization of the two-group nonexperimental design described in Section 2.2.

As explained in Section 2.2, the population means are interpreted differently in experimental and nonexperimental designs. In a nonexperimental design, μ_j describes the mean of the response variable for the subpopulation of people who belong to level j of the classification factor. Experimental designs that include both a treatment factor and a classification factor are described in Sections 3.8 and 3.12.

3.3 Linear Contrasts

Some research questions can be expressed in terms of a *linear contrast* of population means, $\sum_{j=1}^{a} v_j \mu_j$, where v_j is called a *contrast coefficient*. For example, in an experiment that compares two costly treatments (Treatment 1 and Treatment 2) with a new inexpensive treatment (Treatment 3), a confidence interval for $(\mu_1 + \mu_2)/2 - \mu_3$ may provide valuable information regarding the relative costs and benefits of the new treatment. This linear contrast can be expressed as $(.5)\mu_1 + (.5)\mu_2 + (-1)\mu_3$ where the contrast coefficients are $v_1 = .5$, $v_2 = .5$, and $v_3 = -1$. Statistical

packages that have options to compute a confidence interval for a linear contrast of means require the user to specify the contrast coefficients. Consider another example where Treatment 1 is delivered to groups 1 and 2 by experimenters E_1 and E_2 and Treatment 2 is delivered to groups 3 and 4 by experimenters E_3 and E_4. The linear contrast $(\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4)/2$ can be expressed as $(½)\mu_1 + (½)\mu_2 + (-½)\mu_3 + (-½)\mu_4$ and the contrast coefficients are $v_1 = .5$, $v_2 = .5$, $v_3 = -.5$, and $v_4 = -.5$. Note that linear contrast coefficients always sum to 0.

A $100(1-\alpha)\%$ confidence interval for $\sum_{j=1}^{a} v_j \mu_j$ is

$$\sum_{j=1}^{a} v_j \hat{\mu}_j \pm t_{\alpha/2;df} SE_{\sum_{j=1}^{a} v_j \hat{\mu}_j} \tag{3.1}$$

where $df = \left[\sum_{j=1}^{a} \frac{v_j^2 \hat{\sigma}_j^2}{n_j}\right]^2 / [\sum_{j=1}^{a} \frac{v_j^4 \hat{\sigma}_j^4}{n_j^2 (n_j - 1)}]$ and $SE_{\sum_{j=1}^{a} v_j \hat{\mu}_j} = \sqrt{\sum_{j=1}^{a} \frac{v_j^2 \hat{\sigma}_j^2}{n_j}}$ is the standard error of the estimated linear contrast.

The v_j coefficients in Equation 3.1 are not required to sum to 0. If the v_j coefficients do not sum to 0, then $\sum_{j=1}^{a} v_j \mu_j$ is referred to as a *linear function* of means and the coefficients will be referred to as *weights*. Examples of linear functions of means are given in section 3.17.

If the sample sizes are approximately equal, the standard error in Equation 3.1 could be replaced with an equal-variance standard error of $\sqrt{\hat{\sigma}_p^2 \sum_{j=1}^{a} v_j^2 / n_j}$ where $\hat{\sigma}_p^2 = [\sum_{j=1}^{a} (n_j - 1)\, \hat{\sigma}_j^2]/df$ and $df = (\sum_{j=1}^{a} n_j) - a$. The `ci.lc.mean.bs` function in the `statpsych` package computes both versions of Equation 3.1.

If g planned linear contrasts are computed, α in Equation 3.1 can be replaced with $\alpha^* = \alpha/g$ to obtain a set of simultaneous confidence intervals. In an exploratory analysis where no linear contrasts are planned in advance and one or more linear contrasts appear interesting *after* an examination of the sample results, it is necessary to replace $t_{\alpha/2;df}$ in Equation 3.1 with a *Scheffé* critical value $\sqrt{(a-1)F_{\alpha;a-1;df}}$ where $F_{\alpha;a-1;df}$ is a critical F-value (see Appendix A). If equal variances are assumed in Equation 3.1 then $df = (\sum_{j=1}^{a} n_j) - a$; otherwise, the df for a Welch F test (see section 3.20) can be used.

Example 3.2. Ninety students were randomly selected from a research participant pool and randomized into three groups. Then all students listened to an audio recording that listed the names of 40 people who will be attending a party. In group 1, the participants were asked to draw squares while they were listening to the audio recording. In group 2, the participants were not told to draw anything while listening. In group 3, the participants were told to draw copies of complex geometric figures while listening. All three groups were given the same set of boring tasks for 20 minutes and then asked to recall the names of the party attendees. The number of correctly recalled attendees was obtained from each participant. The sample means and variances are given below.

Squares	No Drawing	Complex Drawing
$\hat{\mu}_1 = 24.9$	$\hat{\mu}_2 = 23.1$	$\hat{\mu}_3 = 16.4$
$\hat{\sigma}_1 = 5.21$	$\hat{\sigma}_2 = 4.67$	$\hat{\sigma}_3 = 4.98$
$n_1 = 30$	$n_2 = 30$	$n_3 = 30$

The 95% confidence interval for $(\mu_1 + \mu_2)/2 - \mu_3$ is [5.38, 9.82]. The researcher is 95% confident that the population mean number of correctly recalled attendees averaged across the no drawing and simple drawing conditions is 5.38 to 9.82 greater than the population mean correctly recalled attendees under the complex drawing condition.

3.4 Standardized Linear Contrasts

In applications where the scale of the response variable might be unfamiliar to the intended audience, it could be helpful to report a confidence interval for a *standardized linear contrast* of population means. The following standardized linear contrast is appropriate for experimental designs and is a generalization of Equation 2.2.

$$\varphi = \frac{\sum_{j=1}^{a} v_j \mu_j}{\sqrt{(\sum_{j=1}^{a} \sigma_j^2)/a}} \tag{3.2}$$

An approximate $100(1-\alpha)\%$ confidence interval for φ is

$$\hat{\varphi} \pm z_{\alpha/2} SE_{\hat{\varphi}} \tag{3.3}$$

where $\hat{\varphi} = \sum_{j=1}^{a} v_j \hat{\mu}_j / \hat{\sigma}$, $SE_{\hat{\varphi}} = \sqrt{(\hat{\varphi}^2/2a^2)\sum_{j=1}^{a} \frac{\hat{\sigma}_j^4}{\hat{\sigma}^4(n_j-1)} + \sum_{j=1}^{a} \frac{v_j^2 \hat{\sigma}_j^2}{\hat{\sigma}^2(n_j-1)}}$, and $\hat{\sigma} = \sqrt{(\sum_{j=1}^{a} \hat{\sigma}_j^2)/a}$. Note that $\hat{\sigma}$ is the square root of an *unweighted* average of variances.

The following standardized linear contrast is appropriate for nonexperimental designs with simple random sampling and is a generalization of Equation 2.4.

$$\varphi' = \frac{\sum_{j=1}^{a} v_j \mu_j}{\sqrt{\sum_{j=1}^{a} \pi_j \sigma_j^2}} \tag{3.4}$$

where π_j is the proportion of the study population that belongs to level j of the classification factor. An approximate $100(1-\alpha)\%$ confidence interval for φ' is

$$\widehat{\varphi'} \pm z_{\alpha/2} SE_{\hat{\varphi}} \tag{3.5}$$

where $\widehat{\varphi'} = \sum_{j=1}^{a} v_j \hat{\mu}_j / \hat{\sigma}_p$ and $SE_{\hat{\varphi}} = \sqrt{(\hat{\varphi}^2/2a^2)\sum_{j=1}^{a} \frac{1}{(n_j - 1)} + \sum_{j=1}^{a} \frac{v_j^2 \hat{\sigma}_j^2}{\hat{\sigma}_p^2 n_j}}$. Note that $\hat{\sigma}_p$ is the square root of a *weighted* average of variances. If the population variances are approximately equal, Equation 3.5 is also appropriate for nonexperimental designs with stratified random sampling or for experimental designs.

A third type of standardized linear contrast, denoted as φ'', replaces $\sqrt{(\sum_{j=1}^{a} \sigma_j^2)/a}$ in Equation 3.2 with the standard deviation in one group (usually a control group). This standardized linear contrast is appropriate in both experimental and nonexperimental designs. The interpretation of φ and φ' are most clear and useful when the variances are similar. The interpretability of φ'' is not affected by unequal variances.

The `ci.lc.stdmean.bs` function in the `statpsych` package will compute a confidence interval for all three types of standardized linear contrasts using a method that does not assume equal population variances. SPSS can compute a confidence interval for φ' using a method that assumes equal population variances.

Estimates of standardized linear contrasts have a slight positive bias in small samples. The bias can be reduced by multiplying $\hat{\varphi}$ and $\hat{\varphi}'$ by $1 - 3/[4(\sum_{j=1}^{a} n_j) - 4a - 1]$ and multiplying $\hat{\varphi}$" by $1 - 3/[4n_j - 5]$ where n_j is the sample size of the group used to specify the standardizer. Note that the bias adjustments are not needed or recommended in the computation of the confidence intervals.

Example 3.3. One hundred and sixty students were randomly selected from a research participant pool of about 2,000 students and randomly divided into four group of equal size. In each group, participants were told to study a picture for 30 seconds and be ready to answer questions about the objects in the picture. Each group viewed a different picture. Group 1 viewed a picture of a car with five men in the background, group 2 viewed a picture of a car with five women in the background, group 3 viewed a picture of a car with five dogs in the background, and group 4 viewed a picture of a car with five cats in the background. An eye tracker was used to measure the amount of eye movement during the 30-second viewing period. The sample means and standard deviations are given below.

Group 1	Group 2	Group 3	Group 4
$\hat{\mu}_1 = 6.94$	$\hat{\mu}_2 = 7.15$	$\hat{\mu}_3 = 4.60$	$\hat{\mu}_4 = 3.68$
$\hat{\sigma}_1 = 2.21$	$\hat{\sigma}_2 = 2.83$	$\hat{\sigma}_3 = 2.29$	$\hat{\sigma}_4 = 1.90$

The researcher is interested in assessing the linear contrast $(\mu_1 + \mu_2)/2 - (\mu_3 - \mu_4)/2$. The estimated standardized linear contrast is 1.24 and the 95% confidence interval is [0.90, 1.59]. In the population of 2,000 students, the researcher is 95% confident that the average of the two eye-movement population means for pictures with men and women in the background is 0.90 to 1.59 standard deviations greater than the average of the two eye-movement population means for pictures with dogs and cats in the background.

3.5 Simultaneous Directional Two-sided Tests

Simultaneous confidence intervals could be used to test multiple hypotheses and keep the *familywise directional error rate* (FWDER) at or below $\alpha/2$. FWDER is the probability of making one or more directional errors when testing multiple null hypotheses. The *Holm* method can be used to perform multiple planned hypothesis tests. The Holm method keeps the FWDER at or below $\alpha/2$ and is more powerful than tests based on simultaneous confidence intervals.

To implement the Holm method for g planned null hypotheses, rank order the p-values for the g tests from smallest to largest. If the smallest p-value is less than α/g, then reject H_0 for that test and examine the next smallest p-value; otherwise, do not reject H_0 for that test or *any* of the remaining $g - 1$ null hypotheses. If the second smallest p-value is less than $\alpha/(g - 1)$, then reject H_0 for that test and examine the next smallest p-value; otherwise, do not reject H_0 or *any* of the remaining $g - 2$ null hypotheses. If the third smallest p-value is less than $\alpha/(g - 2)$, then reject H_0 for that test and examine the next smallest p-value; otherwise, do not reject H_0 or *any* of the remaining $g - 3$ null hypotheses (and so on). Suppose the

ranked p-values for three ($g = 3$) tests of linear contrasts are .004, .028, and .031. For $\alpha = .05$, the first null hypothesis is rejected because .004 < .05/3. The second null hypothesis is not rejected because .028 > .05/2. The third null hypothesis also is not rejected because the second null hypothesis was not rejected even though .031 < .05.

The Bonferroni method also could be used to perform multiple directional two-sided tests while keeping the FWDER at or below $\alpha/2$. To obtain Bonferroni simultaneous tests, every p-value is compared with α/g. Note that the Holm test is more powerful than the Bonferroni test because with the Holm test only the smallest p-value is compared with α/g and the other p-values are compared with larger adjusted alpha values.

3.6 Hypothesis Tests for Linear Contrasts

A confidence interval for $\sum_{j=1}^{a} v_j \mu_j$ can be used to perform a directional two-sided test of the following hypotheses.

$$H_0: \sum_{j=1}^{a} v_j \mu_j = 0 \qquad H_1: \sum_{j=1}^{a} v_j \mu_j > 0 \qquad H_2: \sum_{j=1}^{a} v_j \mu_j < 0$$

If the lower limit for $\sum_{j=1}^{a} v_j \mu_j$ is greater than 0, then reject H_0 and accept H_1. If the upper limit for $\sum_{j=1}^{a} v_j \mu_j$ is less than 0, then reject H_0 and accept H_2. The results are inconclusive if the confidence interval includes 0. Note that it is not necessary to develop special hypothesis testing rules for φ because $\sum_{j=1}^{a} v_j \mu_j = 0$ implies $\varphi = 0$, $\sum_{j=1}^{a} v_j \mu_j > 0$ implies $\varphi > 0$, and $\sum_{j=1}^{a} v_j \mu_j < 0$ implies $\varphi < 0$. The test statistic for a linear contrast is $t = \sum_{j=1}^{a} v_j \hat{\mu}_j / SE_{\sum_{j=1}^{a} v_j \mu_j}$. SPSS will compute t and its p-value for a test of H_0: $\sum_{j=1}^{a} v_j \mu_j = 0$ using the equal-variance method and also the unequal-variance method. A p-value less than .05 does not imply that $\sum_{j=1}^{a} v_j \mu_j$ is a meaningfully large value, and a p-value greater than .05 does not imply that $\sum_{j=1}^{a} v_j \mu_j = 0$.

If g tests of linear contrasts are planned, the Holm test can be used to keep the FWDER at or below $\alpha/2$. In an exploratory analysis where the linear contrasts are not planned and one or more linear contrasts appear interesting after an examination of the sample results, it is necessary to compare the test statistic with the *Scheffé* critical value described in Section 3.3.

3.7 One-way Analysis of Variance

The variability in the response variable scores in a one-factor design can be decomposed into two sources of variability – the variance of scores within treatments (called the *error variance* or *residual variance*) and the variance due to mean differences across treatments (called *between-group variance*). The decomposition of variability in a one-factor design can be summarized in a one-way analysis of variance (one-way ANOVA) table, as shown below, where n is the total sample size ($n = n_1 + n_2 + \ldots + n_a$), SS stands for *sum of squares,* and MS stands for *mean square*. The between-group factor (i.e., the independent variable) will be referred to as "Factor A". The ANOVA table for a one-factor design is shown below.

Source	SS	df	MS	F
A	SS_A	$df_A = a - 1$	$MS_A = SS_A/df_A$	MS_A/MS_E
ERROR	SS_E	$df_E = n - a$	$MS_E = SS_E/df_E$	
TOTAL	SS_T	$df_T = n - 1$		

The sum of squares formulas are given below.

$$SS_A = \sum_{j=1}^{a} n_j \left(\hat{\mu}_j - \hat{\mu}_+\right)^2 \text{ where } \hat{\mu}_+ = \sum_{j=1}^{a}\sum_{i=1}^{n_j} y_{ij} / \sum_{j=1}^{a} n_j \tag{3.6}$$

$$SS_E = \sum_{j=1}^{a}\sum_{i=1}^{n_j} \left(y_{ij} - \hat{\mu}_j\right)^2 = \sum_{j=1}^{a} \left(n_j - 1\right) \hat{\sigma}_j^2 \tag{3.7}$$

$$SS_T = \sum_{j=1}^{a}\sum_{i=1}^{n_j} \left(y_{ij} - \hat{\mu}_+\right)^2 = SS_A + SS_E \tag{3.8}$$

SS_A will equal zero if all sample means are equal and will be large if the sample means are highly unequal. $MS_E = SS_E/df_E$ is called the *mean squared error* and is equal to the pooled within-group variance ($\hat{\sigma}_p^2$) that was defined previously.

The SS values in the ANOVA table can be used to estimate a standardized measure of effect size called *eta-squared*. The population value of eta-squared can be defined as $\eta^2 = 1 - \sigma_E^2/\sigma_T^2$. In a nonexperimental design, σ_T^2 is the variance of the response variable for everyone in the study population, and σ_E^2 is the variance of the response variable within each subpopulation of the study population. In an experimental design, σ_E^2 is the variance of the response variable for every person

in the study population assuming they all received a particular treatment and $\sigma_T^2 = \sigma_\mu^2 + \sigma_E^2$ where σ_μ^2 is the variance of the population means ($\mu_1, \mu_2, \ldots, \mu_a$) under the *a* treatment conditions. Unlike pairwise comparisons or linear contrasts, η^2 does not provide any information about how the population means differ.

An estimate of η^2 can be computed using any of the following four formulas

$$\hat{\eta}^2 = 1 - \frac{SS_E}{SS_T} = \frac{SS_A}{SS_A + SS_E} = \frac{SS_A}{SS_T} = \frac{df_A \, x \, F}{df_A \, x \, F + df_E} \tag{3.9}$$

where $F = MS_A/MS_E$. The value of $\hat{\eta}^2$ can range from 0 to 1 (because SS_E has a possible range of 0 to SS_T) and describes the proportion of the response variable variance in the sample that is predictable from the between-group factor. The estimate of η^2 in Equation 3.9 tends to overstate the value of η^2. The following *adjusted eta-squared* is less biased

$$adj\ \hat{\eta}^2 = 1 - (df_A + df_E)(1 - \hat{\eta}^2)/df_E \tag{3.10}$$

and should be reported along with a confidence interval for η^2. SPSS and the `ci.etasqr` function in the `statpsych` package will compute a confidence interval for η^2.

Example 3.4. Sixty undergraduates were randomly selected from a study population of 4,350 college students and then classified into three groups according to their political affiliation (Democrat, Republican, Independent). A stereotyping questionnaire was given to all 60 participants. A one-way ANOVA in this nonexperimental design detected differences in the three population means ($F(2, 57) = 5.02$, $p = .010$, $\hat{\eta}^2 = .15$, $adj\ \hat{\eta}^2 = .12$, 95% CI [.01, .30]). The researcher can be 95% confident that 1% to 30% of the variance in the stereotyping scores of the 4,350 college students can be predicted from knowledge of their political affiliation. This study needs to be replicated with a larger sample size because the lower limit for η^2 suggests that the effect of political affiliation could be trivial while the upper limit suggests that political affiliation could be very important.

The *F* statistic from the ANOVA table is used to test the null hypothesis H_0: $\mu_1 = \mu_2 = \ldots = \mu_a$ against an alternative hypothesis that at least one pair of population means is not equal. This type of hypothesis test is referred to as an *omnibus test*. Rejecting H_0: $\mu_1 = \mu_2 = \ldots = \mu_a$ does not reveal anything about how the population means are ordered or the magnitudes of the population mean differences. If $a = 2$, then the *F* statistic is equal to the squared equal-variance *t*

statistic defined in section 2.7. Statistical packages will compute the *p*-value for the *F* statistic that is used to decide if H_0 can be rejected. A *p*-value less than .05 does not imply that there is an important difference in the population means, and a *p*-value greater than .05 does not imply that the population means are equal. In one-factor studies where $a > 2$ and the *p*-value is less than .05, a common mistake is to assume that the order of the population means corresponds to the order of the sample means.

The rejection of H_0: $\mu_1 = \mu_2 = \ldots = \mu_a$ is not a scientifically important finding because this null hypothesis is known to be false in almost every study. Furthermore, a "nonsignificant" result should not be interpreted as evidence that the null hypothesis is true. However, the omnibus test is sometimes useful in exploratory studies. If the null hypothesis cannot be rejected, then no further exploratory analyses are required because every possible $100(1-\alpha)\%$ linear contrast confidence interval using a Scheffé critical value will include 0 if the *p*-value for the omnibus test is greater than α.

Some researchers conduct a preliminary test of H_0: $\mu_1 = \mu_2 = \ldots = \mu_a$ and then follow-up with tests or confidence intervals of pairwise comparisons or linear contrasts only if H_0: $\mu_1 = \mu_2 = \ldots = \mu_a$ can be rejected. This preliminary test approach is not required or recommended when using the Holm, Bonferroni, or Scheffé methods. However, in the special case of $a = 3$, the preliminary test approach is useful. If H_0: $\mu_1 = \mu_2 = \mu_3$ is rejected, then no correction to the α value is needed for any follow-up hypothesis tests of pairwise comparisons or linear contrasts to keep the FWDER at or below $\alpha/2$. These follow-up tests are called *Fisher protected* tests and can be more powerful than tests based on the Holm method.

The omnibus null and alternative hypotheses can also be expressed as H_0: $\eta^2 = 0$ and H_1: $\eta^2 > 0$, respectively. If the lower limit of a $100(1-\alpha)\%$ *one-sided* lower confidence interval for η^2 is greater than 0, then H_0: $\eta^2 = 0$ is rejected. Note that if the *p*-value for the *F* statistic is less than α, then the one-sided lower limit must be greater than 0.

A two-sided confidence interval for η^2 can be used to test H_1: $\eta^2 \leq h$ against H_2: $\eta^2 > h$ where *h* is a value, such as .02, that would be considered a "small" effect size. Accept H_2 if the lower limit is greater than *h*; accept H_1 if the upper limit is

less than h; otherwise, the results are inconclusive. Accepting H_1 is evidence that the population means ($\mu_1, \ \mu_2, \ \ldots \ , \ \mu_a$) all have similar values. Accepting H_2 is evidence that there are nontrivial differences among the population means. Note that a two-sided $100(1-\alpha)\%$ confidence interval for η^2 will have a lower limit greater than 0 only of the p-value for the F statistic is less than $\alpha/2$.

3.8 Two-Factor Designs

In a one-factor experiment, the researcher is able to assess the causal effect of only one independent variable on the response variable. The effect of two independent variables on the response variable can be assessed in a *two-factor experiment*. The two factors will be referred to as "Factor A" and "Factor B". The simplest type of two-factor experiment has two levels of Factor A and two levels of Factor B. We call this a 2×2 factorial experiment. If Factor A had 4 levels and Factor B had 3 levels, it would be called a 4×3 factorial experiment. In general, an $a \times b$ factorial experiment has a levels of Factor A and b levels of Factor B.

There are three types of two-factor between-subjects experiments. In one case, both factors are treatment factors and participants are randomly assigned to the combinations of treatment conditions. A second type of two-factor experiment uses one treatment factor and one classification factor where stratified random sampling has been used to sample from the levels of the classification factor. In this type of experiment, a random sample of participants is obtained from each level of the classification factor and are then randomly assigned to the treatment conditions within each level of the classification factor. A third type of two-factor experiment uses one treatment factor and one classification factor where simple random sampling is used to obtain a single sample of participants. The participants are first classified into the levels of the classification factor and are then randomly assigned to the treatment conditions within each level of the classification factor. A study could also have two classification factors, but then it would be a two-factor nonexperimental design. Some examples of two-factor experiments are given below.

Example 3.5. An experiment with two treatment factors takes randomly sampled low-income students and randomly assigns them to one of four treatment conditions: 1) faculty mentor and a student learning community, 2) graduate student mentor and a student learning community, 3) faculty mentor and no student learning community, and 4) graduate student mentor and no student learning community. One treatment factor is the type of mentor (faculty or graduate student) and the other treatment factor is participation or nonparticipation in a student learning community. The response variable is an academic self-efficacy questionnaire.

Example 3.6. An experiment with one classification factor and one treatment factor uses a stratified random sample of 30 men and 30 women from a volunteer list of students taking introductory psychology. The samples of men and women are each randomly assigned into two groups with one group receiving 2 hours and the other group receiving 6 hours of supplemental instruction with a graduate student instructor. The treatment factor is the amount of review (2 or 6 hours) and the classification factor is gender. The response variable is the final exam score.

Example 3.7. An experiment with one classification factor and one treatment factor obtained a simple random sample of undergraduate college students and classified them into first-generation and continuing-generation groups. Then each group of students was randomly assigned to receive mentoring from a graduate student or a faculty member. The response variable is a college life satisfaction questionnaire.

One advantage of a two-factor experiment is that the effects of both Factor A and Factor B can be assessed in a single study. Questions about the effects of Factor A and Factor B could be answered using two separate one-factor experiments. However, two one-factor experiments would require at least twice the total number of participants to obtain confidence intervals with the same precision or hypothesis tests with the same power that could be obtained from a single two-factor experiment. Thus, a two-factor experiment is more economical than two one-factor experiments.

A two-factor experiment can also provide information that cannot be obtained from two one-factor experiments. Specifically, a two-factor experiment can provide unique information about the *interaction effect* between Factor A and Factor B. An interaction effect occurs when the effect of Factor A differs across the levels of Factor B or effect of Factor B differs across the levels of Factor A. A more precise definition of an interaction effect is given in section 3.9

Adding a second factor can improve the external validity of an experiment. For example, if there is a concern that participants might perform a particular task differently in the morning than in the afternoon, then time of day (e.g., morning vs.. afternoon) could be used as a second 2-level factor in the experiment. If the interaction effect between Factor A and the time-of-day factor (Factor B) is small, then the effect of Factor A would generalize to both morning and afternoon testing conditions, thus increasing the external validity of the results for Factor A.

The external validity of an experiment can also be improved by including a classification factor. If the interaction between the classification factor and the treatment factor is small, then the effect of the treatment factor can be generalized to the multiple study populations that are represented by the levels of the classification factor, thereby increasing the external validity of the results for the treatment factor.

Adding a classification factor to an experiment can reduce the error variance (MS_E), which will increase the power of statistical tests and reduce the widths of confidence intervals. For example, in a one-factor experiment with male and female subjects, if women tend to score higher than men, this will increase the variance of scores within treatments. If gender is added as a classification factor, the error variance will then be determined by the variability of scores within each treatment and within each gender, which will result in a smaller MS_E.

3.9 Definition of Effects in Two-Factor Designs

Consider the special case of a 2 × 2 factorial design. The population means for this design are shown below.

		Factor B	
		b_1	b_2
Factor A	a_1	μ_{11}	μ_{12}
	a_2	μ_{21}	μ_{22}

The *main effects* of Factor A and Factor B and the AB *interaction effect* are defined below.

A: $(\mu_{11} + \mu_{12})/2 - (\mu_{21} + \mu_{22})/2$

B: $(\mu_{11} + \mu_{21})/2 - (\mu_{12} + \mu_{22})/2$

AB: $(\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22}) = (\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22})$

The *simple main effects* of A and B are defined below.

A at b_1: $\mu_{11} - \mu_{21}$ B at a_1: $\mu_{11} - \mu_{12}$

A at b_2: $\mu_{12} - \mu_{22}$ B at a_2: $\mu_{21} - \mu_{22}$

The interaction effect can be expressed as a difference in simple main effects, specifically: $(\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22})$ = (B at a_1) – (B at a_2), or equivalently, $(\mu_{11} - \mu_{21}) - (\mu_{12} - \mu_{22})$ = (A at b_1) – (A at b_2). The main effects can be expressed as averages of simple main effects. The main effect of A is (A at b_1 + A at b_2)/2 = $(\mu_{11} - \mu_{21} + \mu_{12} - \mu_{22})/2 = (\mu_{11} + \mu_{12})/2 - (\mu_{21} + \mu_{22})/2$. The main effect of B is (B at a_1 + B at a_2)/2 = $(\mu_{11} - \mu_{12} + \mu_{21} - \mu_{22})/2 = (\mu_{11} + \mu_{21})/2 - (\mu_{12} + \mu_{22})/2$.

All of the above effects are special cases of a linear contrast of means, and confidence intervals for these effects can be obtained using Equation 3.1 by specifying the appropriate contrast coefficients. Alternatively, the `ci.2x2.mean.bs` function in the `statpsych` package will compute confidence intervals for the AB interaction, the two main effects, and the four simple main effects in a 2 × 2 design.

If the AB interaction effect is large, the main effect of A (which is the average of A at b_1 and A at b_2) could be misleading because A at b_1 and A at b_2 will be highly dissimilar. Likewise, the main effect of B (which is the average of B at a_1 and B at a_2) could be misleading if the AB interaction is large because B at a_1 and B at a_2 will be highly dissimilar. If the AB interaction effect is large, then an analysis of simple main effects will be more meaningful than an analysis of main effects. If the AB interaction is small, then the main effects will not be misleading and an analysis of simple main effects will be unnecessary.

3.10 Pairwise Main Effects and Simple Main Effects

In a two-factor experiment where a factor has more than two levels, various pairwise comparisons can be examined. Consider a 2 × 3 design where the main effects of Factor B are of interest. The population means are given below.

		Factor B		
		b_1	b_2	b_3
Factor A	a_1	μ_{11}	μ_{12}	μ_{13}
	a_2	μ_{21}	μ_{22}	μ_{23}

The following three *pairwise main effects* can be defined for Factor B

B_{12}: $(\mu_{11} + \mu_{21})/2 - (\mu_{12} + \mu_{22})/2$
B_{13}: $(\mu_{11} + \mu_{21})/2 - (\mu_{13} + \mu_{23})/2$
B_{23}: $(\mu_{12} + \mu_{22})/2 - (\mu_{13} + \mu_{23})/2$

where the subscripts of B represent the levels of the factor being compared.

If any factor has more than two levels, then more than one interaction effect can be defined. An interaction effect can be defined for any two levels of Factor A and any two levels of Factor B. For example, in the 2 × 3 design described above, the following three *pairwise interaction effects* can be defined

$A_{12}B_{12}$: $(\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22})$
$A_{12}B_{13}$: $(\mu_{11} - \mu_{13}) - (\mu_{21} - \mu_{23})$
$A_{12}B_{23}$: $(\mu_{12} - \mu_{13}) - (\mu_{22} - \mu_{23})$

where the subscripts of AB represent the levels of Factor A and Factor B being compared. The number of pairwise interaction effects can be overwhelming in larger designs. For example, in a 4 × 3 design, there are six pairs of Factor A levels and three pairs of Factor B levels from which 6 × 3 = 18 pairwise interaction effects could be examined.

If any factor has more than two levels and the AB interaction effect is large, *pairwise simple main effects* will provide useful information. In the 2 × 3 design described above, Factor B has three levels and the pairwise simple main effects of Factor B are defined below.

B_{12} at a_1: $\mu_{11} - \mu_{12}$ B_{12} at a_2: $\mu_{21} - \mu_{22}$
B_{13} at a_1: $\mu_{11} - \mu_{13}$ B_{13} at a_2: $\mu_{21} - \mu_{23}$
B_{23} at a_1: $\mu_{12} - \mu_{13}$ B_{23} at a_2: $\mu_{22} - \mu_{23}$

3.11 Main Effect and Simple Main Effect Linear Contrasts

If any factor in a two-factor design has three or more levels, a *main effect linear contrast* or *simple main effect linear contrast* could be more interesting than main effect or simple main effect pairwise comparisons. Suppose the 4-level factor in a 2 × 4 design has the following levels: b_1 = teaching method 1 with instructor 1, b_2 = teaching method 1 with instructor 2, b_3 = teaching method 2 with instructor 3, and b_4 = teaching method 2 with instructor 4. The population means for this 2 × 4 design are given below.

		Factor B			
		b_1	b_2	b_3	b_4
Factor A	a_1	μ_{11}	μ_{12}	μ_{13}	μ_{14}
	a_2	μ_{21}	μ_{22}	μ_{23}	μ_{24}

In this study the researcher is most interested in comparing teaching method 1 with teaching method 2. Suppose an analysis of the AB interaction suggests that main effects should be examined. The main effect linear contrast to compare the two teaching methods is $(\frac{\mu_{11}+\mu_{21}}{2}+\frac{\mu_{12}+\mu_{22}}{2})/2 - (\frac{\mu_{13}+\mu_{23}}{2}+\frac{\mu_{14}+\mu_{24}}{2})/2$.

Suppose an analysis of the AB interaction suggests that simple main effects should be examined. The simple linear contrasts to compare the two teaching methods are $(\mu_{11} + \mu_{12})/2 - (\mu_{13} + \mu_{14})/2$ at a_1 and $(\mu_{21} + \mu_{22})/2 - (\mu_{23} + \mu_{24})/2$ at a_2.

Confidence intervals for all of the effects in Section 3.10 and this section can be expressed as $\sum_j^{ab} v_j \mu_j$ where *ab* is the total number of groups and can be computed using Equation 3.1. For example, the contrast coefficients that define $(\frac{\mu_{11}+\mu_{21}}{2}+\frac{\mu_{12}+\mu_{22}}{2})/2 - (\frac{\mu_{13}+\mu_{23}}{2}+\frac{\mu_{14}+\mu_{24}}{2})/2$ (assuming the means in the 2 × 4 table are ordered left to right and then top to bottom) are 1/4, 1/4, -1/4, -1/4, 1/4, 1/4, -1/4, and -1/4. A confidence interval for $\sum_j^{ab} v_j \mu_j$ can also be used to test a variety of hypotheses regarding the value of $\sum_j^{ab} v_j \mu_j$.

3.12 Two-Way Analysis of Variance

Now consider a general $a \times b$ factorial design. The variability of the response variable scores in a two-factor design can be decomposed into four sources of variability: the variance due to differences in means across the levels of Factor A, the variance due to differences in means across the levels of Factor B, the variance due to differences in simple main effects of one factor across the levels of the other factor (the AB interaction), and the variance of scores within treatments (the error variance). The decomposition of the total variance in a two-factor design can be summarized in the following two-way analysis of variance (two-way ANOVA) table where n is the total sample size.

Source	*SS*	*df*	*MS*	*F*
A	SS_A	$df_A = a - 1$	$MS_A = SS_A/df_A$	MS_A/MS_E
B	SS_B	$df_B = b - 1$	$MS_B = SS_B/df_B$	MS_B/MS_E
AB	SS_{AB}	$df_{AB} = (a-1)(b-1)$	$MS_{AB} = SS_{AB}/df_{AB}$	MS_{AB}/MS_E
ERROR	SS_E	$df_E = n - ab$	$MS_E = SS_E/df_E$	
TOTAL	SS_T	$df_T = n - 1$		

The TOTAL and ERROR sum of squares formulas in a two-way ANOVA shown below are conceptually similar to the one-way ANOVA formulas

$$SS_T = \sum_{k=1}^{b}\sum_{j=1}^{a}\sum_{i=1}^{n_{jk}}(y_{ijk} - \hat{\mu}_{++})^2 \tag{3.11}$$

$$SS_E = \sum_{k=1}^{b}\sum_{j=1}^{a}\sum_{i=1}^{n_{jk}}(y_{ijk} - \hat{\mu}_{jk})^2 \tag{3.12}$$

where $\hat{\mu}_{++} = (\sum_{k=1}^{b}\sum_{j=1}^{a}\sum_{i=1}^{n_{jk}} y_{ijk})/(\sum_{k=1}^{b}\sum_{j=1}^{a} n_{jk})$ is the mean of all y scores ignoring group membership. The formulas for SS_A, SS_B, and SS_{AB} are complicated unless the sample sizes are equal. If all sample sizes are equal to n_0, the formulas for SS_A, SS_B, and SS_{AB} simplify as shown below.

$$SS_A = bn_0\sum_{j=1}^{a}\left(\hat{\mu}_{j+} - \hat{\mu}_{++}\right)^2 \quad \text{where } \hat{\mu}_{j+} = \sum_{k=1}^{b}\sum_{i=1}^{n_0} y_{ij}/bn_0 \tag{3.13}$$

$$SS_B = an_0\sum_{k=1}^{b}(\hat{\mu}_{+k} - \hat{\mu}_{++})^2 \quad \text{where } \hat{\mu}_{+k} = \sum_{j=1}^{a}\sum_{i=1}^{n_0} y_{ij}/an_0 \tag{3.14}$$

$$SS_{AB} = SS_T - SS_E - SS_A - SS_B. \tag{3.15}$$

If both factors are treatment factors, *partial eta-squared* estimates can be computed from the sum of squares estimates as shown below.

$$\hat{\eta}^2_{\text{A}} = SS_{\text{A}}/(SS_{\text{T}} - SS_{\text{B}} - SS_{\text{AB}}) = SS_{\text{A}}/(SS_{\text{A}} + SS_{\text{E}}) \quad (3.16a)$$

$$\hat{\eta}^2_{\text{B}} = SS_{\text{B}}/(SS_{\text{T}} - SS_{\text{A}} - SS_{\text{AB}}) = SS_{\text{B}}/(SS_{\text{B}} + SS_{\text{E}}) \quad (3.16b)$$

$$\hat{\eta}^2_{\text{AB}} = SS_{\text{AB}}/(SS_{\text{T}} - SS_{\text{B}} - SS_{\text{A}}) = SS_{\text{AB}}/(SS_{\text{AB}} + SS_{\text{E}}) \quad (3.16c)$$

These measures are called "partial" effect sizes because the denominator for a particular factor removes the effects of all other factors. For example, SS_{B} and SS_{AB} are subtracted from SS_{T} to obtain $\hat{\eta}^2_{A}$, and SS_{A} and SS_{AB} are subtracted from SS_{T} to obtain $\hat{\eta}^2_{B}$. SPSS and R have options to compute these partial eta-square estimates.

Now suppose Factor A is a treatment factor and Factor B is a classification factor. The following *generalized eta-squared* measures of effect size are recommended.

$$\hat{\eta}^2_{\text{A}} = SS_{\text{A}}/(SS_{\text{A}} + SS_{\text{B}} + SS_{\text{AB}} + SS_{\text{E}}) = SS_{\text{A}}/SS_{\text{T}} \quad (3.17a)$$

$$\hat{\eta}^2_{\text{B}} = SS_{\text{B}}/(SS_{\text{B}} + SS_{\text{AB}} + SS_{\text{E}}) \quad (3.17b)$$

$$\hat{\eta}^2_{\text{AB}} = SS_{\text{AB}}/(SS_{\text{AB}} + SS_{\text{B}} + SS_{\text{E}}) \quad (3.17c)$$

Generalized eta-squared estimates are recommended in factorial designs that have a classification factor because the *SS* for the classification factor and its interaction with the treatment factor describes natural variation among participants that should not be removed from SS_{T}. If both factors are classification factors (a nonexperimental design), the recommended generalized eta-squared estimates are

$$\hat{\eta}^2_{\text{A}} = SS_{\text{A}}/(SS_{\text{A}} + SS_{\text{B}} + SS_{\text{AB}} + SS_{\text{E}}) = SS_{\text{A}}/SS_{\text{T}} \quad (3.18a)$$

$$\hat{\eta}^2_{\text{B}} = SS_{\text{B}}/(SS_{\text{A}} + SS_{\text{B}} + SS_{\text{AB}} + SS_{\text{E}}) = SS_{\text{B}}/SS_{\text{T}} \quad (3.18b)$$

$$\hat{\eta}^2_{\text{AB}} = SS_{\text{AB}}/(SS_{\text{A}} + SS_{\text{B}} + SS_{\text{AB}} + SS_{\text{E}}) = SS_{\text{AB}}/SS_{\text{T}}. \quad (3.18c)$$

All of the above eta-squared estimates are positively biased. The adjusted eta-square estimate for a single-factor design (Equation 3.10), is given below for any effect and reduces the bias of both partial and generalized eta-squared estimates.

$$adj\ \hat{\eta}^2_{\text{effect}} = 1 - (df_{\text{effect}} + df_{\text{E}})(1 - \hat{\eta}^2_{\text{effect}})/\ df_{\text{E}} \quad (3.19)$$

In the special case of a single-factor design, Equation 3.10 is also referred to as "epsilon-squared". It is important to accompany an eta-squared estimate (preferably the adjusted estimate) with a confidence interval for the population eta-squared value. A confidence interval for a population partial or generalized eta-squared value can be obtained using the `ci.etasqr` function in the `statpsych` package.

The *F* statistics for the main effect of Factor A, the main effect of Factor B, and the AB interaction effect, test null hypotheses regarding population eta-squared values: H_0: $\eta_A^2 = 0$, H_0: $\eta_B^2 = 0$, and H_0: $\eta_{AB}^2 = 0$. Tests of these omnibus null hypotheses suffer from the same problem as the test of the omnibus null hypothesis in a one-way ANOVA. Specifically, a "significant" result does not indicate that a scientifically important result has been obtained, and a "nonsignificant" result does not imply that the effect is zero. The *F* statistics and *p*-values for each effect should be supplemented with confidence intervals for population eta-squared values, linear contrasts of population means, or linear contrasts of unstandardized population means.

3.13 Analysis Strategies for Two-factor Designs

An examination of the AB interaction is usually the first step in the analysis of a two-factor design. The AB interaction can be assessed using the *p*-value for the AB interaction in a two-way ANOVA along with a confidence interval for η_{AB}^2. If the *p*-value for the test of the AB interaction effect is less than .05, it is customary to only analyze simple main effects or pairwise simple main effects. However, a main effect could be interesting when the *p*-value for the AB interaction is less than .05 if the main effect is substantially larger than the interaction effect as assessed by the appropriate eta-squared estimates.

Although an inconclusive result for the test of the AB interaction effect does not imply that the population interaction effect is zero, it is customary to examine main effects rather than simple main effects if the AB interaction test is inconclusive. However, if the study was specifically designed to assess the direction and magnitude of simple main effects, then simple main effects should be examined even if the test for the interaction effect is inconclusive.

Directional two-sided hypothesis tests for simple main effects can also be used to justify an analysis of main effects, even if the *p*-value for the interaction effect is less than .05. For example, if the directional tests for the simple main effects of Factor A all lead to the same directional conclusions at different levels of Factor B, then the main effect of Factor A could be examined. A less stringent justification allows some of the simple main effect tests to be inconclusive as long as the estimated simple main effects for a particular factor all have the same sign (e.g., the simple main effect estimates of Factor A are all positive or are all negative).

3.14 Three-factor Designs

The effects of three independent variables can be assessed in a three-factor design. The three factors will be referred to as "Factor A", "Factor B", and "Factor C". Like a two-factor design, a three-factor design provides information about main effects and two-way interaction effects. The main effects of Factors A, B, and C can be estimated as well as the AB, AC, and BC two-way interactions. These main effects and two-way interaction effects could be estimated from three separate two-factor studies. A three-factor study has the advantage of providing all this information in a single study and also providing information about a *three-way interaction* (ABC) that could not be obtained from separate two-factor studies. The factors in a three-factor design can be treatment factors or classification factors. In an experimental design, at least one of the three factors is a treatment factor.

The simplest type of three-factor design has two levels of each factor and is called a 2 × 2 × 2 factorial design. In general, $a \times b \times c$ factorial designs have a levels of Factor A, b levels of Factor B, and c levels of Factor C. A table of population means is shown below for a 2 × 2 × 2 factorial design.

		Factor C			
		c_1		c_2	
		Factor B		Factor B	
		b_1	b_2	b_1	b_2
Factor A	a_1	μ_{111}	μ_{121}	μ_{112}	μ_{122}
	a_2	μ_{211}	μ_{221}	μ_{212}	μ_{222}

In this $2 \times 2 \times 2$ factorial design, the main effects of Factors A, B, and C are defined as,

A: $(\mu_{111} + \mu_{121} + \mu_{112} + \mu_{122})/4 - (\mu_{211} + \mu_{221} + \mu_{212} + \mu_{222})/4$
B: $(\mu_{111} + \mu_{211} + \mu_{112} + \mu_{212})/4 - (\mu_{121} + \mu_{221} + \mu_{122} + \mu_{222})/4$
C: $(\mu_{111} + \mu_{211} + \mu_{121} + \mu_{221})/4 - (\mu_{112} + \mu_{212} + \mu_{122} + \mu_{222})/4,$

the three two-way interaction effects are defined as,

AB: $(\mu_{111} + \mu_{112})/2 - (\mu_{121} + \mu_{122})/2 - (\mu_{211} + \mu_{212})/2 + (\mu_{221} + \mu_{222})/2$
AC: $(\mu_{111} + \mu_{121})/2 - (\mu_{112} + \mu_{122})/2 - (\mu_{211} + \mu_{221})/2 + (\mu_{212} + \mu_{222})/2$
BC: $(\mu_{111} + \mu_{211})/2 - (\mu_{112} + \mu_{212})/2 - (\mu_{121} + \mu_{221})/2 + (\mu_{122} + \mu_{222})/2,$

and the three-way interaction effect is defined as

ABC: $\mu_{111} - \mu_{121} - \mu_{211} + \mu_{221} - \mu_{112} + \mu_{122} + \mu_{212} - \mu_{222}.$

The *simple main effects* of Factors A, B, and C are defined below.

A at b_1: $(\mu_{111} + \mu_{112})/2 - (\mu_{211} + \mu_{212})/2$
A at b_2: $(\mu_{121} + \mu_{122})/2 - (\mu_{221} + \mu_{222})/2$
A at c_1: $(\mu_{111} + \mu_{121})/2 - (\mu_{211} + \mu_{221})/2$
A at c_2: $(\mu_{112} + \mu_{122})/2 - (\mu_{212} + \mu_{222})/2$

B at a_1: $(\mu_{111} + \mu_{112})/2 - (\mu_{121} + \mu_{122})/2$
B at a_2: $(\mu_{211} + \mu_{212})/2 - (\mu_{221} + \mu_{222})/2$
B at c_1: $(\mu_{111} + \mu_{211})/2 - (\mu_{121} + \mu_{221})/2$
B at c_2: $(\mu_{112} + \mu_{212})/2 - (\mu_{122} + \mu_{222})/2$

C at a_1: $(\mu_{111} + \mu_{121})/2 - (\mu_{112} + \mu_{122})/2$
C at a_1: $(\mu_{211} + \mu_{221})/2 - (\mu_{212} + \mu_{222})/2$
C at b_1: $(\mu_{111} + \mu_{211})/2 - (\mu_{112} + \mu_{212})/2$
C at b_2: $(\mu_{121} + \mu_{221})/2 - (\mu_{122} + \mu_{222})/2$

The *simple-simple main effects* of Factors A, B, and C are defined as,

A at b_1c_1: $\mu_{111} - \mu_{211}$
A at b_1c_2: $\mu_{112} - \mu_{212}$
A at b_2c_1: $\mu_{121} - \mu_{221}$
A at b_2c_2: $\mu_{122} - \mu_{222}$

B at a_1c_1: $\mu_{111} - \mu_{121}$
B at a_1c_2: $\mu_{112} - \mu_{122}$
B at a_2c_1: $\mu_{211} - \mu_{221}$
B at a_2c_2: $\mu_{212} - \mu_{222}$

C at a_1b_1: $\mu_{111} - \mu_{112}$
C at a_1b_2: $\mu_{121} - \mu_{122}$
C at a_2b_1: $\mu_{211} - \mu_{212}$
C at a_2b_2: $\mu_{221} - \mu_{222},$

and the *simple two-way interaction effects* are defined as

$$\begin{aligned}
&\text{AB at } c_1\text{: } (\mu_{111} - \mu_{121}) - (\mu_{211} - \mu_{221}) \quad \text{AB at } c_2\text{: } (\mu_{112} - \mu_{122}) - (\mu_{212} - \mu_{222})\\
&\text{AC at } b_1\text{: } (\mu_{111} - \mu_{211}) - (\mu_{112} - \mu_{212}) \quad \text{AC at } b_2\text{: } (\mu_{121} - \mu_{221}) - (\mu_{122} - \mu_{222})\\
&\text{BC at } a_1\text{: } (\mu_{111} - \mu_{121}) - (\mu_{112} - \mu_{122}) \quad \text{BC at } a_2\text{: } (\mu_{211} - \mu_{221}) - (\mu_{212} - \mu_{222}).
\end{aligned}$$

The ABC interaction in a $2 \times 2 \times 2$ design can be defined as a difference in simple two-way interaction effects. Specifically, the ABC interaction is the difference between AB at c_1 and AB at c_2, or the difference between AC at b_1 and AC at b_2, or the difference between BC at a_1 and BC at a_2. Although the meaning of a three-way interaction is not easy to grasp, its meaning becomes clearer when it is viewed as the difference in simple two-way interaction effects with each simple two-way interaction interpreted as a difference in simple-simple main effects.

The two-way interaction effects in a three-factor design are conceptually the same as in a two-factor design. Two-way interactions in a three-factor design are defined by collapsing the three-dimensional table of population means to create a two-dimensional table of means with cell means that have been averaged over the collapsed dimension. For example, a table of averaged population means in a $2 \times 2 \times 2$ factorial design, after collapsing Factor C, gives the following 2×2 table from which the AB interaction can be defined in terms of the averaged population means.

		Factor B	
		b_1	b_2
Factor A	a_1	$(\mu_{111} + \mu_{112})/2$	$(\mu_{121} + \mu_{122})/2$
	a_2	$(\mu_{211} + \mu_{212})/2$	$(\mu_{221} + \mu_{222})/2$

3.15 Three-Way Analysis of Variance

The variability of the response variable in a three-factor design can be decomposed into eight sources of variability: three main effects, three two-way interactions, one three-way interaction, and the within-group error variance. The decomposition of the total variance in a three-factor design can be summarized in the following

three-way analysis of variance (three-way ANOVA) table where n is the total sample size.

Source	SS	df	MS	F
A	SS_A	$df_A = a - 1$	$MS_A = SS_A/df_A$	MS_A/MS_E
B	SS_B	$df_B = b - 1$	$MS_B = SS_B/df_B$	MS_B/MS_E
C	SS_C	$df_C = c - 1$	$MS_C = SS_C/df_C$	MS_C/MS_E
AB	SS_{AB}	$df_{AB} = (a - 1)(b - 1)$	$MS_{AB} = SS_{AB}/df_{AB}$	MS_{AB}/MS_E
AC	SS_{AC}	$df_{AC} = (a - 1)(c - 1)$	$MS_{AC} = SS_{AC}/df_{AC}$	MS_{AC}/MS_E
BC	SS_{BC}	$df_{BC} = (b - 1)(c - 1)$	$MS_{BC} = SS_{BC}/df_{BC}$	MS_{BC}/MS_E
ABC	SS_{ABC}	$df_{ABC} = df_A df_B df_C$	$MS_{ABC} = SS_{ABC}/df_{ABC}$	MS_{ABC}/MS_E
ERROR	SS_E	$df_E = n - abc$	$MS_E = SS_E/df_E$	
TOTAL	SS_T	$df_T = n - 1$		

A partial eta-squared estimate can be computed for each of the seven effects in a three-way ANOVA. These partial eta-squared estimates are computed the same way they are computed in a two-way ANOVA. For example, $\hat{\eta}_A^2 = SS_A/(SS_A + SS_E)$ and $\hat{\eta}_{ABC}^2 = SS_{ABC}/(SS_{ABC} + SS_E)$. If any of the factors are classification factors, then generalized eta-squared estimates are usually recommended. The denominators of the generalized eta-squared estimates include the SS estimates for all main effects and interaction effects that involve a classification factor. For example, if Factor C is a classification factor, then $\hat{\eta}_A^2 = SS_A/(SS_A + SS_C + SS_{AC} + SS_{BC} + SS_{ABC} + SS_E)$. Equation 3.19 can be used to reduce the positive bias in partial and generalized eta-squared estimates in three-factor designs.

The seven omnibus F tests in the three-way ANOVA suffer from the same problem as the omnibus F tests in the one-way and two-way ANOVA. These tests should be supplemented with confidence intervals for population eta-squared values, linear contrast of population means, or standardized linear contrasts of population means to provide information regarding the magnitude of each effect.

3.16 Analysis Strategies for Three-factor Designs

The results from a three-way ANOVA can be used to determine a variety of different follow-up analyses which are summarized below.

- If the p-value for the ABC test is less than .05, then simple two-way interactions or simple-simple main effects should be examined.

- If the ABC test is inconclusive, then the AB, AC, and BC interactions should be examined.

- If the ABC test is inconclusive and if *all three* two-way interactions have p-values less than .05, then simple-simple main effects should be examined.

- If the ABC test is inconclusive and only *two* of the two-way interactions have p-values less than .05, then the appropriate simple main effects should be examined. For example, if the BC interaction is the one inconclusive interaction, then it would not be necessary to examine the simple main effects of B at c_1 and c_2 or the simple main effects of C at b_1 and b_2.

- If the ABC test is inconclusive only *one* of the two-way interactions has a p-value less than .05, then simple main effects for the two interacting factors should be examined and the main effect for the third factor should be examined.

- If the tests for the ABC, AB, AC, and BC interactions are all inconclusive, then the main effects of A, B, and C should be examined.

Although it is customary to base the above rules on p-values, a modification of these rules can be expressed in terms of eta-squared estimates. For example, if the p-value for the ABC test is less than .05 but the estimate of η^2_{AB} is substantially larger than the estimate of η^2_{ABC}, then it might be appropriate to examine the AB interaction rather than the simple AB interactions. Likewise, an analysis of a main effect can be justified even if that factor interacts with another factor but the eta-squared estimate for the main effect is substantially larger than the interaction effect.

Directional two-sided tests of simple effects can also be used to determine the appropriate effect to analyze. As one example, if the simple-simple main effects of

Factor A at each combination of the Factor B and Factor C levels lead to the same directional conclusions, it could be appropriate to examine the main effect of Factor A even if the *p*-values for any of the ABC, AB, BC, or AC interactions are less than .05. A less stringent justification would only require the simple-simple main effect estimates of Factor A to all have the same sign.

If a study is specifically designed to assess simple main effects or simple-simple main effects because of their theoretical or practical importance, these simple effects can be analyzed regardless of the preliminary interaction test results. This alternative analysis strategy can be justified in terms of sample size requirements because the sample size required to test an interaction effect with desired power can be substantially larger than the sample size required to test a simple main effect or a simple-simple main effect (see section 3.23).

3.17 Subpopulation Size Weighting

If a study includes a classification factor, the means at each level of the classification factor can be weighted by known subpopulation sizes to define a mean for the total population. Consider a two-group nonexperimental design that obtained a random sample of 20 first-generation students, another random sample of 20 continuing-generation students, and students in both samples were given a campus safety opinion questionnaire. The main purpose of the study was to compare the means of the first-generation and continuing-generation students, but a secondary purpose was to estimate the mean opinion score in the total population of first-generation and continuing-generation students. Suppose the researcher knows that 21% of all students on campus are first-generation students and 79% are continuing-education students. The total population mean can then be defined as $.21\mu_1 + .79\mu_2$, and Equation 3.1 can be used to compute a confidence interval for this linear function of means.

Subpopulation size weighting can also be used in factorial designs that have one or more classification factors. Consider a 2 × 3 factorial design with a 2-level treatment factor (e.g., in-person instruction vs. remote instruction) and a 3-level classification factor (e.g., Caucasian, Hispanic, and Other). Suppose an analysis of the interaction effect suggests that the main effect of treatment should be examined. Referring to the 2 × 3 table of population means in section 3.10, the

traditional main effect of treatment is defined as $(\mu_{11} + \mu_{12} + \mu_{13})/3 - (\mu_{21} + \mu_{22} + \mu_{23})/3$. This definition gives equal importance to all three ethnicities which could be appropriate in certain applications. Alternatively, the main effect of treatment in the total population requires subpopulation size weighting. Suppose the proportions of Caucasian, Hispanic, and Other students on campus are .68, .21, and .11, respectively. The main effect of treatment for the total population is then defined as $.68(\mu_{11} - \mu_{21}) + .21(\mu_{12} - \mu_{22}) + .11(\mu_{13} - \mu_{23}) = .68\mu_{11} + .21\mu_{12} + .11\mu_{13} - .68\mu_{21} - .21\mu_{22} - .11\mu_{23}$, and Equation 3.1 can be used to compute a confidence interval for this linear function of means.

Example 3.8. A random sample of 60 first-year female social science students and a random sample of 60 first-year male social science students were randomly assigned to learning community housing for social science students (Treatment 1) or regular housing (Treatment 2). GPA at the end of the first year was determined for all 120 students. The test for the interaction effect was inconclusive, and it was decided to examine the main effect of housing type for all first-year social science students. At this university, the proportion of first-year female social science students is .65 and the proportion of first-year male social science students is .35. The main effect of housing type (Treatment 1 vs. Treatment 2) for all first-year social science students is $.65\mu_{11} + .35\mu_{12} - .65\mu_{21} - .35\mu_{22}$, and the 95% confidence interval for this linear function of means is [0.25, 0.41]. This result suggests that if all first-year social students could live in a social science learning community, their mean GPA at the end of the first year would be 0.25 to 0.41 higher than if they had instead all lived in regular housing.

3.18 One-way Random Effects ANOVA

All of the factors considered up to this point have been *fixed factors* because it was assumed that the factor levels used in the study were deliberately selected and were the only factor levels of interest. In comparison, the levels of a *random factor* are randomly selected from a population of *M* possible factor levels. The appeal of using a random factor is that the statistical results apply to all *M* levels of the random factor even though only a small subset of the factor levels were used in the study.

Recall that a factor can be a classification factor or a treatment factor. Recall also that the levels of a classification factor define different subpopulations. Although a random factor can be a classification factor or a treatment factor, most random

factors in psychology are classification factors, and only random classification factors will be illustrated here.

In studies where *M* subpopulations could be examined, such as all schools in a state, all neighborhoods in a large city, or all branch offices of a large organization, it could be costly or impractical to take a random sample of participants from each of the *M* subpopulations. In these situations, the researcher could select a random sample of subpopulations from the set of *M* subpopulations and then take a random sample of n_j participants from each of the subpopulations. This type of sampling is called *two-stage cluster sampling*.

In a *one-way random effects ANOVA,* the subpopulation means of interest are μ_1, μ_2, … , μ_M. The *grand mean* is $\mu = (\mu_1 + \mu_2 + \ldots + \mu_M)/M$, and the standard deviation of the *M* subpopulation means is $\sigma_\mu = \sqrt{\sum_{j=1}^{M}(\mu_j - \mu)^2 / M}$. A one-way random effects ANOVA with $a \geq 2$ levels of the random factor provides estimates of μ_1, μ_2, … , μ_a where these subpopulation means are a random sample from the *M* subpopulation means. In the one-way random effects ANOVA, σ_μ is a measure of effect size because larger values of σ_μ indicate larger differences among the *M* subpopulation means. A standardized measure of effect size in a one-way random effects ANOVA is $\omega^2 = \sigma_\mu^2/(\sigma_\mu^2 + \sigma_e^2)$ where σ_e^2 is the within-subpopulation variance.

Like η^2 for nonexperimental designs, ω^2 describes the proportion of variance of the y scores in the total population that is explained by differences in the subpopulation means. Although η^2 and ω^2 have the same basic interpretation in nonexperimental designs, different symbols are used because the confidence intervals for η^2 and ω^2 are different. Also, ω^2 describes the proportion of explained variance for all *M* subpopulations while η^2 describes the proportion of explained variance just in those subpopulations that were used in the study.

The one-way random effects ANOVA table is identical to the one-way (fixed effects) ANOVA table described in Section 3.7. If the sample sizes in each of the randomly selected subpopulations are equal, then the parameter estimates and confidence interval formulas are greatly simplified.

Although only $a < M$ population means are estimated in a study that uses a random classification factor, it is possible to obtain a confidence interval for $\mu = (\mu_1 + \mu_2 + \ \ldots \ + \mu_M)/M$. With equal sample sizes per group, a $100(1-\alpha)\%$ confidence interval for μ is

$$\hat{\mu} \pm t_{\alpha/2;(a-1)}\sqrt{MS_A/an_0} \tag{3.20}$$

where n_0 is the sample size per group, $\hat{\mu} = (\hat{\mu}_1 + \hat{\mu}_2 + \ \ldots \ + \hat{\mu}_a)/a$, and MS_A is the mean square estimate for the one-way fixed effects ANOVA given in Section 3.7.

The MS estimates in the one-way fixed-effects ANOVA table can be used to estimate σ_e^2, σ_μ, and ω^2 as shown below.

$$\hat{\sigma}_e^2 = MS_E \tag{3.21}$$

$$\hat{\sigma}_\mu^2 = (MS_A - MS_E)/n_0 \tag{3.22}$$

$$\hat{\omega}^2 = (MS_A - MS_E)/[MS_A + (n_0 - 1)MS_E] \tag{3.23}$$

Unlike $\hat{\eta}^2$, $\hat{\omega}^2$ is nearly unbiased and does not require a bias adjustment.

An approximate $100(1-\alpha)\%$ confidence interval for σ_e is

$$\sqrt{exp[ln(\hat{\sigma}_e^2) \pm z_{\alpha/2}SE_{\hat{\sigma}_e^2}/\hat{\sigma}_e^2]} \tag{3.24}$$

and an approximate confidence interval for σ_μ is

$$\sqrt{exp[ln(\hat{\sigma}_\mu^2) \pm z_{\alpha/2}SE_{\hat{\sigma}_\mu^2}/\hat{\sigma}_\mu^2]} \tag{3.25}$$

where $SE_{\hat{\sigma}_e^2} = \sqrt{\frac{2MS_E^2}{a(n_0-1)}}$ and $SE_{\hat{\sigma}_\mu^2} = \sqrt{\frac{2MS_E^2}{n_0^2 a(n_0-1)} + \frac{2MS_A^2}{n_0^2(a-1)}}$. SPSS and R will compute versions of Equations 3.20, 3.24, and 3.25 that do not require equal sample sizes.

The $100(1-\alpha)\%$ lower and upper confidence limits for ω^2 are

$$LL = (F/F_{\alpha/2;\,df_1,df_2} - 1)/(n_0 + F/F_{\alpha/2;\,df_1,df_2} - 1) \tag{3.26a}$$

$$UL = (F/F_{1-\alpha/2;\,df_1,df_2} - 1)/(n_0 + F/F_{1-\alpha/2;\,df_1,df_2} - 1) \tag{3.26b}$$

where $F_{\alpha/2;\, df_1, df_2}$ and $F_{1-\alpha/2;\, df_1, df_2}$ are critical F values (see Table 4 in Appendix A) with $df_1 = a - 1$ and $df_2 = a(n_0 - 1)$ and $F = MS_A/MS_E$. The `ci.random.anova1` function in the `statpsych` package will compute Equations 3.20, 3.24, 3.25, and 3.26ab for studies with equal sample sizes.

The width of the confidence interval for ω^2 is affected primarily by the number of factor levels and not the sample size per group. This confidence interval will be substantially narrower if, for example, a random sample of $n_0 = 5$ participants is obtained from $a = 30$ randomly selected subpopulations rather than a random sample of $n_0 = 30$ participants from $a = 5$ randomly selected subpopulations.

Example 3.9. A researcher obtained a random sample of 50 elementary schools from a set of about 9,000 elementary schools in California. A random sample of 30 students was obtained from each of the 50 randomly selected schools and a food insecurity questionnaire (scored 0 to 20) was given to the 1,500 randomly selected students. A 95% confidence interval for the grand mean food insecurity score is [14.82, 16.15]. We can be 95% confident that the average of the 9,000 elementary school food insecurity means is between 14.82 and 16.15. A 95% confidence interval for ω^2 is [.215, .432]. The researcher can be 95% confident that 21.5% to 43.2% of the total variance in student food insecurity scores is explained by differences in the subpopulation means of the 9,000 schools.

The null hypothesis in a one-way random effects ANOVA is H_0: $\mu_1 = \mu_2 = \ldots = \mu_M$ which can be expressed as H_0: $\sigma_\mu = 0$. The test statistic is $F = MS_A/MS_E$, which is the same test statistic used in a one-way fixed-effects ANOVA. Rejecting H_0: $\sigma_\mu = 0$ does not provide useful scientific information because σ_μ will almost never exactly equal 0. If H_0: $\sigma_\mu = 0$ is rejected, some researchers will want to conduct an exploratory analysis to search for characteristics of the selected subpopulations that might be related to the response variable, but this type of exploratory analysis assumes that the factor is fixed rather than random. Confidence intervals for linear contrasts of population means using a Scheffé critical value can be used to perform this type of exploratory analysis, but the results will apply only to the selected subpopulations and not the M subpopulations.

3.19 Two-factor Design with a Random Classification Factor

As explained in Section 3.8, adding a fixed classification factor to an experiment will increase the generality of the results. For example, random samples of children could be obtained at three different elementary schools and then randomly assigned to different treatments within each school. If the interaction between the school and treatment factors is small, then the main effect of the treatment factor can be generalized to the study populations of all three schools. The use of a random classification factor can provide even greater levels of generalization. For example, if a random sample of elementary schools is sampled from a large set of M elementary schools and students are randomly assigned to treatments within each of the randomly selected schools, then the main effect of treatment will generalize to all M study populations if the interaction effect is small.

The SS and MS estimates in the two-way ANOVA table are computed in exactly the same way with a fixed classification factor or a random classification factor. However, the F statistic for the main effect of treatment is computed differently. Suppose Factor A is the treatment factor and Factor B is the random classification factor. Recall from Section 3.8, where both factors are fixed, that the F statistic for the main effect of Factor A is $F = MS_A/MS_E$ with $a - 1$ and $n - a$ degrees of freedom. If Factor B is a random classification factor, then the F statistic for the main effect of Factor A is $F = MS_A/MS_{AB}$ with $a - 1$ and $(a - 1)(b - 1)$ degrees of freedom. The F statistic for the AB interaction is computed the same way for a fixed or random classification factor.

There is some controversy regarding the computation of the F statistic for the main effect of the random classification factor (Factor B). The majority of statistics texts recommend using $F = MS_B/MS_E$ but the GLM procedure in SPSS will compute $F = MS_B/MS_{AB}$. The arguments favoring the use of $F = MS_B/MS_E$ are compelling and this test statistic is recommended here. The controversy is much less of an issue when the primary goal of the study is to assess the main effect of the fixed treatment factor. Furthermore, the test for Factor B is an omnibus test of a null hypothesis that the population means are identical in all M study populations, and this null hypothesis will be false in virtually all applications.

Linear contrasts for the fixed treatment main effect can be obtained using Equation 3.1 with the equal-variance standard error. The variance estimate ($\hat{\sigma}_p^2$) is replaced with MS_{AB}, and the *df* for the critical *t*-value set to $(a-1)(b-1)$. The confidence interval results for these linear contrasts apply to all *M* study populations.

The widths of the confidence intervals for linear contrasts of the fixed treatment are affected primarily by the number of levels of the random classification factor and not the sample size per group.

The generalized eta-square estimate for Factor A in a two-factor design with fixed factors is also recommended for a two-factor design where Factor A is fixed and Factor B is a random classification factor. The positive bias in this estimate can be reduced using Equation 3.19.

3.20 Assumptions

The ANOVA *F* tests, the equal-variance Tukey-Kramer confidence intervals for pairwise comparison, and the equal-variance confidence interval for $\sum_{j=1}^{a} v_j \mu_j$ all assume equality of population variances across treatment conditions in addition to the random sampling, independence, and normality assumptions. Violating the equal variance assumption is serious if the sample sizes are unequal. The unequal-variance methods for pairwise comparisons, linear contrasts, and standardized linear contrasts are preferred to the equal-variance methods if the sample sizes are not similar.

The *Welch F test* is a useful alternative to the one-way ANOVA *F* test that relaxes the equal variance assumption. The Welch *F* test is especially useful in applications where the sample sizes are unequal. In the special case of $a = 3$, the Welch *F* test can be used in place of the one-way ANOVA *F*-test when performing Fisher protected tests. SPSS and R have options to compute the Welch *F* test.

The adverse effects of violating the normality assumption on the *F* tests, tests and confidence intervals for pairwise comparisons, and tests and confidence intervals for $\sum_{j=1}^{a} v_j \mu_j$ are usually not serious unless the response variable is highly skewed and the sample size per group is small ($n_j < 20$).

Violating the normality assumption is more serious for confidence intervals of η^2 and standardized linear contrasts. The true coverage probability of a 95% confidence interval for η^2 or a standardized linear contrast can be much smaller than .95 if the response variable is nonnormal within groups. Furthermore, the adverse effects of nonnormality on these confidence intervals is not diminished in large sample sizes. Data transformations are sometimes helpful in reducing within-group nonnormality.

In addition to the assumptions for a one-way fixed effects ANOVA, the one-way random effects ANOVA also assumes that the distribution of the *M* subpopulation means is approximately normal. A 95% confidence interval for the mean of subpopulation means can have a true coverage probability less than .95 if the distribution of subpopulation means is highly skewed and the number of factor levels is small. Increasing the number of random factor levels will reduce the negative effects of skewed subpopulation means. The true coverage probability of a 95% confidence interval for σ_e , σ_μ, and ω^2 can be much smaller than .95 if the distribution of subpopulation means is leptokurtic. Increasing the sample size or the number of random factor levels will not diminish the negative effects of leptokurtosis on the confidence intervals for σ_e, σ_μ, and ω^2.

To assess the degree of within-group nonnormality in a design with $a \geq 2$ groups, subtract $\hat{\mu}_j$ from all of the group j scores then estimate the skewness and kurtosis coefficients from these $n_1 + n_2 + \cdots + n_a$ deviation scores. If the deviation scores are skewed, it might be possible to reduce the skewness by transforming (e.g., log, square root, reciprocal) the response variable scores.

3.21 Distribution-free Methods

If the response variable is skewed, a confidence interval for a linear contrast of population medians may be more appropriate and meaningful than a confidence interval for a linear contrast of population means. An approximate $100(1-\alpha)\%$ confidence interval for $\sum_{j=1}^{a} v_j\theta_j$ is

$$\sum_{j=1}^{a} v_j\hat{\theta}_j \pm z_{\alpha/2}\sqrt{\sum_{j=1}^{a} v_j^2 SE_{\hat{\theta}_j}^2} \tag{3.27}$$

where $SE^2_{\hat{\theta}_j}$ is defined in Equation 1.10. This confidence interval does not assume normality or equality of variances, but it does assume that tied scores are rare. Equation 3.27 can be used to perform a directional two-sided hypothesis test regarding the value of $\sum_{j=1}^{a} v_j \theta_j$. The `ci.lc.median.bs` function in the `statpsych` package will compute Equation 3.27.

The *Kruskal-Wallis test* is a distribution-free alternative to the *F* test in a one-way ANOVA. The Kruskal-Wallis test suffers from the same problem as the *F* test because the null hypothesis is known to be false in virtually every study, and accepting the alternative hypothesis does not provide any information about how the population medians are ordered or the magnitudes of the population median differences. In designs with more than two groups, useful information can be obtained by performing multiple Mann-Whitney tests for some or all pairwise comparisons using the Holm procedure. Simultaneous confidence intervals for pairwise differences or ratios of medians, the Mann-Whitney parameter (π) for pairwise comparisons, or linear contrasts of medians are informative alternatives to the Kruskal-Wallis test. In the special case of $a = 3$, the Kruskal-Wallis test can be used to obtain Fisher protected tests of distribution-free pairwise comparisons and linear contrasts.

3.22 Multiple Confidence Intervals and Hypothesis Tests in Factorial Designs

Methods for computing simultaneous confidence intervals or controlling the FWDER is fairly straightforward in one-factor designs, but the problem is far more complicated and controversial in factorial designs. Some practical suggestions are given here.

In a two-factor design, test the AB interaction using an unadjusted α value (usually .05) to decide if main effects or simple main effects should be examined. Then use an adjusted α value to analyze main effects or simple main effects. For example, suppose the test for AB interaction in a 2 × 4 design is inconclusive and the two main effects will be examined. There is one main effect for the 2-level factor and 4(3)/2 = 6 pairwise main effect comparisons for the 4-level factor. Simultaneous confidence intervals for these seven main effects can be computed using $\alpha^* = \alpha/7$.

Now suppose the p-value for the AB interaction in the 2×4 design is less than .05 and simple main effects will be examined. The researcher could compute simultaneous confidence intervals for the four simple main effects of Factor A using $\alpha^* = \alpha/4$ or the six pairwise comparisons for Factor B at each of the two levels of Factor A using $\alpha^* = \alpha/12$. If both factors are treatment factors and either set of simple main effects provides a useful description of the nature of the interaction effect, then the four simple main effects of Factor A would be preferred to an analysis of the 12 simple pairwise main effects of Factor B because using $\alpha^* = \alpha/4$ rather than $\alpha^* = \alpha/12$ will give narrower confidence intervals and more powerful tests. If the researcher is interested only in hypothesis testing for the simple main effects or main effects, then the Holm test could be used where the effect with the smallest p-value would be tested using α^*.

Methods for computing simultaneous confidence intervals or controlling the FWDER in three-factor designs follow the same general approach as in two-factor designs. Use an unadjusted α value to test the three-way and two-way interaction effects to determine if main effects, simple main effects, or simple-simple main effects should be examined. Then determine how many effects need to be examined and divide the α value by this number.

Consider a $2 \times 2 \times 2$ design and suppose the interaction tests suggest that simple-simple main effects should be examined. If the researcher decides that the simple-simple main effects of A at b_1c_1, b_2c_1, b_1c_2, and b_2c_2 are the most interesting, then set $\alpha^* = \alpha/4$. If the interaction tests had instead suggested that the three main effects should be examined, then set $\alpha^* = \alpha/3$.

An additional adjustment to the α value is recommended if r different response variables are analyzed. With r response variables and a set of g effects to examine, the adjusted α value is $\alpha^* = \alpha/(rg)$.

A smaller α^* value decreases the power of the hypothesis tests and increases the widths of the confidence intervals. Alternatively, a smaller α^* value will increase the sample size requirement to achieve the desired hypothesis testing power or desired confidence interval precision. In the following sections on sample size planning, α can be replaced with α^* to achieve the desired widths of simultaneous confidence intervals or to test multiple directional two-sided hypotheses while

maintaining the FWDER. However, if α^* is very small, the sample size requirement can be prohibitively large and the researcher will want to carefully assess the need for every proposed response variable, every proposed factor, and every proposed factor level.

3.23 Using Prior Information

A published study might report a "significant" effect in a two-group experiment, but the reported confidence interval for a difference in population means might be so wide that it is not possible to determine if the effect is trivial or perhaps very important. If the published study is replicated in a new study population, the results from the prior study can be compared and combined with the results of the follow-up study. The follow-up study would use the same variables that were used in the prior study, and the follow-up study would be carried out in the same way as the prior study. The sample size in the follow-up study should be at least as large as the sample size in the prior study.

If the sample means, sample standard deviations, and sample sizes have been reported in the prior study, this information can be used to analyze a 2 × 2 design where the two levels of Factor A are the two treatment conditions that were used in the prior study and the two levels of Factor B represent the prior study and the follow-up study as illustrated below.

		Factor B	
		Prior	Follow-up
Factor A	a_1	μ_{11}	μ_{12}
	a_2	μ_{21}	μ_{22}

A confidence interval for the AB interaction effect provides evidence of replication (see Chapter 4 of Volume 5 for more details). If the test for the AB interaction effect is inconclusive, then the main effect of Factor A could be examined. A confidence interval for the AB interaction and main effect of Factor A can be obtained using Equation 3.1 and can be computed using the `ci.lc.mean.bs` function.

A confidence interval for the main effect of Factor A can be substantially narrower than the confidence interval for the mean difference that was reported in the prior study. Furthermore, the results for the main effect of Factor A describe the study populations for both the prior study and the follow-up study.

Example 3.10. An eyewitness identification study with 20 participants per group assessed participants' certainty on a 0 to 10 scale in their selection of the suspect individual from a photo lineup after viewing a short video of a crime scene. Two treatment conditions were assessed. In the first treatment condition, the participants were told that the suspect "will be" in a 5-person photo lineup. In the second treatment condition, participants were told that the suspect "might be" in a 5-person photo lineup. The suspect was included in the lineup in both instruction conditions. The estimated means were 7.67 and 6.33 and the estimated standard deviations were 1.72 and 2.35 in the "will be" and "might be" conditions, respectively. The p-value was less than .05 but the 95% confidence interval was [0.02, 2.66] which suggests that the effect could be close to 0. This study was replicated at another university with 40 participants per group. The estimated means are 7.21 and 5.78, and the estimated standard deviations are 1.56 and 2.04 in the "will be" and "might be" conditions, respectively. The test for the interaction effect was inconclusive, and the 95% confidence interval for the main effect of "will be" vs. "might be" was [0.49, 2.02]. We can be 95% confident that the population mean certainty score averaged over the two study populations would be 0.49 to 2.02 points greater with a "will be" instruction than with a "might be" instruction.

Prior results from a factorial design can also be compared and combined with the results from a follow-up study. For example, if an $a \times b$ factorial design is replicated in a follow-up study, the resulting design would be an $a \times b \times 2$ design and would be analyzed using the same strategies that are described in section 3.16.

As explained in section 1.32, the Bayesian approach uses a subjective prior distribution rather than the actual results from a prior study (although results from prior studies will often influence the subjective prior distribution). Equations 1.21 and 1.22 can be applied to a linear contrast of population means by replacing $\tilde{\mu}$ and $\sigma_{\tilde{\mu}}$ with a subjective mean and standard deviation of a prior distribution of plausible values of the population linear contrast. This generalization of Equation 1.21 and 1.22 applies to all of the linear contrasts defined in this chapter for between-subject designs as well as the linear contrasts of population means for within-subjects designs in Chapter 4. The `ci.bayes.normal` function in the `statpsych` package can be used in any application with a normal prior

distribution and where the sampling distribution of the parameter estimates has an approximate normal distribution. Thus, Equations 1.21 and 1.22 can also be modified for other parameters such as a median, median difference, linear contrast of medians, standardized mean difference, or standardized linear contrast of medians.

As noted in section 1.32, subjective prior distributions are usually not well received in scientific research. An exception is the use of a *skeptical* prior distribution that assumes a small effect size. For example, a skeptic might be 95% confident that a population standardized mean difference will be between -.25 and .25 in a typical psychological experiment. In this example, the prior mean could be set to 0 and the prior standard deviation could be set to .25/1.96 = .128. Suppose the sample standardized mean difference in an experiment is .52 with a standard error of .24 (95% CI [.05, .99]). Applying the `ci.bayes.normal` function gives a posterior mean of .12 with a 95% credible interval of [-.11, .34], which includes 0. In this example, the sample size and sample standardized mean difference were not large enough to overcome the initial skepticism. A skeptical prior should not be used in applications where the goal of the study is to show that the effect size is small.

3.24 Sample Size Requirements for Desired Precision

Equation 2.5 for the two-group design can be generalized to linear contrasts of population means. In a one-factor or factorial design with a total of m groups, the sample size requirement per group to estimate a linear contrast of population means with desired confidence and desired confidence interval width is approximately

$$n_j = 4\tilde{\sigma}^2\left(\sum_{j=1}^{m} v_j^2\right)\left(\frac{z_{\alpha/2}}{w}\right)^2 + \frac{z_{\alpha/2}^2}{2m^*} \tag{3.28}$$

where $\tilde{\sigma}^2$ is a planning value of the average within-group variance, w is the desired confidence interval width, and m^* is the number of non-zero v_j values. The MS_E from previous research could be used to specify the planning value for the average within-group variance. The strategies for specifying a variance planning value described in section 1.29 can also be used to specify a within-group variance planning value. The `size.ci.lc.mean.bs` function in the `statpsych` package will compute Equation 3.28.

Example 3.11. In a proposed study of gender ideology, the researcher wants to estimate $(\mu_{11} + \mu_{12})/2 - (\mu_{21} + \mu_{22})/2$ in a 2 × 2 factorial experiment with 95% confidence, a desired confidence interval width of 3.0, and a planning value of 8.0 for the average within-group error variance. The contrast coefficients are 1/2, 1/2, -1/2, and -1/2. The sample size requirement per group is approximately $n_j = 4(8.0)(1/4 + 1/4 + 1/4 + 1/4)(1.96/3.0)^2 + 0.48 = 14.2 \approx 15$.

The sample size requirement per group to estimate a standardized linear contrast of population means (φ) with desired confidence and desired confidence interval width is approximately

$$n_j = [2\tilde{\varphi}^2/m \ + 4(\sum_{j=1}^{m} v_j^2)](\frac{z_{\alpha/2}}{w})^2 \tag{3.29}$$

where $\tilde{\varphi}$ is a planning value of φ or φ' and w is the desired confidence interval width. Equation 3.29 reduces to Equation 2.6 for the special case of a pairwise comparison. The sample size requirement per group to estimate φ" with desired confidence and desired confidence interval width is approximately

$$n_j = [2\tilde{\varphi}"^2 \ + 4(\sum_{j=1}^{m} v_j^2)](\frac{z_{\alpha/2}}{w})^2 \tag{3.30}$$

where $\tilde{\varphi}$" is a planning value of φ". The sample size requirement is larger for $\tilde{\varphi}$" than φ or φ'. The `size.ci.lc.stdmean.bs` function in the `statpsych` package will compute Equations 3.29 and 3.30. Set the planning value of the standardized linear contrast to its largest likely value for a conservatively large sample size requirement. Replace α with $\alpha^* = \alpha/g$ in Equations 3.28 – 3.30 for g simultaneous confidence intervals.

It is often easier to specify the desired width of a confidence interval for φ than a confidence interval for $\sum_{j=1}^{m} v_j\mu_j$. If the researcher plans to report a confidence interval for $\sum_{j=1}^{m} v_j\mu_j$ and finds it easier to specify the desired confidence interval width for φ, then multiplying the desired confidence interval width for φ by $\sqrt{\tilde{\sigma}^2}$ gives the corresponding desired confidence interval width for $\sum_{j=1}^{m} v_j\mu_j$ which can be used in Equation 3.28.

Example 3.12. In a proposed racial stereotype study, a researcher wants to estimate φ in a one-factor experiment ($a = 3$) with 95% confidence, a desired confidence interval width of 0.6, and $\tilde{\varphi} = 0.8$. The contrast coefficients are 1/2, 1/2, and -1. The sample size requirement per group is approximately $n_j = [2(0.64)/3 + 4(1/4 + 1/4 + 1)](1.96/0.6)^2 = 68.6 \approx 69$.

The `size.ci.lc.median.bs` function in the `statpsych` package will compute the sample size required estimate $\sum_{j=1}^{m} v_j\theta_j$ with desired confidence and desired confidence interval width. The required sample size depends on the shape of the distribution of y scores (see section 1.28).

The sample size per group needed to estimate eta-squared with desired confidence and desired confidence interval width is approximately

$$n_j = [16\tilde{\eta}^2(1 - \tilde{\eta}^2)^2(z_{\alpha/2}/w)^2 + m + 1]/m \tag{3.31}$$

where $\tilde{\eta}^2$ is a planning value for the population eta-squared, w is the desired confidence interval width, and m is the total number of groups. The `size.ci.etasqr` function in the `statpsych` package computes Equation 3.31 as a first-step approximation and then applies an adjustment for greater accuracy.

Specifying a narrow confidence interval width in Equations 3.28 - 3.21 might require a prohibitively large sample size. Rather than specifying a desirably narrow confidence interval width, the researcher could also try specifying the largest confidence interval width that would still provide useful information.

Example 3.13. A one-factor nonexperimental design will be used to examine the relation between political affiliation (Democrat, Republican, or Independent) and attitudes towards gun safety in a study population of young working adults. The planning value of η^2 was set at .2 and the desired width of a 95% confidence interval for η^2 was set at .15. The sample size requirement per group is $n_j = [16(.2)(1 - .2)^2(1.96/.15)^2 + 3]/3 = 116.5 \approx 117$. The `size.ci.etasqr` function gives a more accurate sample size of 103 per group.

If sample data can be obtained in two stages, then the confidence interval width for eta-squared obtained in the first-stage sample can be used in Equation 1.18 to approximate the additional number of participants needed in the second-stage sample to achieve the desired confidence interval width. The first-stage and second-stage samples are taken from the same study population, and the participants from the second-stage sample are combined with the participants from the first-stage sample.

Example 3.14. A one-factor experiment with 12 DACA college students per group gave a 95% confidence interval for η^2 with a width of 0.51. The researcher would like to obtain a 95% confidence interval for η^2 that has a width of about 0.25. To achieve this goal, $[(0.51/0.25)^2 - 1]12 = 37.9 \approx 38$ additional DACA students per group are needed.

3.25 Sample Size Requirements for Desired Power

In a one-factor or factorial design with a total of m groups, the sample size requirement per group for a directional two-sided test of H_0: $\sum_{j=1}^{m} v_j \mu_j = 0$ with a specified α value and desired power is approximately

$$n_j = \tilde{\sigma}^2 (\sum_{j=1}^{m} v_j^2)(z_{\alpha/2} + z_\beta)^2 / (\sum_{j=1}^{m} v_j \tilde{\mu}_j)^2 + \frac{z_{\alpha/2}^2}{2m^*} \quad (3.32)$$

where $\tilde{\sigma}^2$ is the planning value of the average within-group variance, $\sum_{j=1}^{m} v_j \tilde{\mu}_j$ is the anticipated effect size value, and m^* is the number of non-zero v_j values. This sample size formula reduces to Equation 2.7 when the contrast involves the comparison of two means. In applications where $\sum_{j=1}^{m} v_j \tilde{\mu}_j$ or $\tilde{\sigma}^2$ is difficult for the researcher to specify, Equation 3.32 can be expressed in terms of a planning value for φ, as shown below

$$n_j = (\sum_{j=1}^{m} v_j^2)(z_{\alpha/2} + z_\beta)^2 / \tilde{\varphi}^2 + \frac{z_{\alpha/2}^2}{2m^*} \quad (3.33)$$

which simplifies to Equation 2.8 when the contrast involves the comparison of two means. The `size.test.lc.mean.bs` in the `statpsych` package function will compute Equation 3.32. This function can also be used to compute Equation 3.33 by setting the variance planning value to 1 and setting the effect size to $\tilde{\varphi}$. Replace α with $\alpha^* = \alpha/g$ in Equations 3.32 and 3.33 for g simultaneous hypothesis tests. For a test of a linear contrast, SPSS has an option to compute the required sample size for the desired power. The strategies for specifying an effect size described in section 1.29 can also be used to specify the effect size in Equations 3.32 and 3.33.

Example 3.15. In a proposed study of retrieval-induced forgetting, a researcher wants to test H_0: $(\mu_1 + \mu_2 + \mu_3 + \mu_4)/4 - \mu_5$ in a one-factor experiment with power of .90, α = .05, and an anticipated standardized linear contrast value of 0.5. The contrast coefficients are 1/4, 1/4, 1/4, 1/4, and -1. The sample size requirement per group is approximately $n_j = 1.25(1.96 + 1.28)^2/0.5^2 + 0.38 = 52.9 \approx 53$.

Several of the above sample size formulas contain the term $\sum_{j=1}^{m} v_j^2$ where a larger value of $\sum_{j=1}^{m} v_j^2$ corresponds to a larger sample size requirement. This has important implications when designing a study with two or more factors. For example, in a 2 × 2 design, $\sum_{j=1}^{m} v_j^2 = 1^2 + (-1)^2 + (-1)^2 + 1^2 = 4$ for the interaction effect, $\sum_{j=1}^{m} v_j^2 = 1^2 + (-1)^2 + 0^2 + 0^2 = 2$ for a simple main effect, and $\sum_{j=1}^{m} v_j^2 = 1/2^2 + 1/2^2 + (-1/2)^2 + (-1/2)^2 = 1$ for a main effect. Note that the sample size requirement for the interaction effect is four times as large as the sample size requirement for a main effect. The sample size required to analyze a two-way or three-way interaction could be prohibitively large, and the researcher might decide to use a sample size that allows main effects, simple main effects, or simple-simple main effects to be estimated with desired precision or tested with desired power.

3.26 Data Transformations and Interaction Effects

Data transformations were described in Chapter 1 as a way to reduce nonnormality. Most psychological measures are assumed to be interval-scale measurements but they might actually be ordinal-scale measurements. Interval-scale measurements are assumed to be linearly related to the attribute they claim to measure while ordinal-scale measurements are assumed to be monotonically related to the attribute. If the y scores have a positive linear relation with the attribute, then every 1-point increase in y will correspond to the same increase in the attribute. But if the y scores have a positive monotonic relation with the attribute, then different 1-point increases in y (e.g., 2 to 3, 8 to 9, etc.) will correspond to different increases in the attribute. A *monotonic transformation* of the y scores (e.g., log, square root, reciprocal) will not change the rank ordering of the y scores. If any monotonic transformation of the y scores substantially reduces the magnitude of an interaction effect, this suggests that the interaction might simply be due to ordinal characteristics of the measurement scale.

Consider the following example of a 2 × 2 design with three participants per group (the sample means are given in parentheses).

		Factor B	
		b_1	b_2
Factor A	a_1	49, 64, 81 ($\hat{\mu}_{11}$ = 64.67)	100, 121, 144 ($\hat{\mu}_{12}$ = 121.67)
	a_2	1, 4, 9 ($\hat{\mu}_{21}$ = 4.67)	16, 25, 36 ($\hat{\mu}_{22}$ = 25.67)

The simple main effect of A at b_1 is 64.67 – 4.67 = 60 and the simple main effect of A at b_2 is 121.67 – 25.67 = 96, which indicates a nonzero interaction effect in this sample. After taking a square root transformation of the data, the sample means are $\hat{\mu}_{11}$ = 8, $\hat{\mu}_{12}$ = 11, $\hat{\mu}_{21}$ = 2, and $\hat{\mu}_{22}$ = 5. After the data transformation, the simple main effect of A at b_1 is 8 – 2 = 6 and the simple main effect of A at b_2 is 11 – 5 = 6, which indicates a zero interaction effect. In this example, the estimated interaction effect was reduced to zero by a monotonic transformation of the data.

Interaction effects can be classified as *removable* or *non-removable*. A removable interaction effect (also called an *ordinal interaction effect*) can be reduced to zero by some monotonic data transformation. A non-removable interaction effect (also called a *disordinal interaction effect*) cannot be reduced to zero by a monotonic data transformation. In a two-factor design, if the simple main effects (or simple pairwise main effects) of Factor A have different signs at different levels of Factor B, *or* the simple main effects (or simple pairwise main effects) of Factor B have different signs at different levels of Factor A, then the interaction effect is non-removable. Otherwise, the interaction effect is potentially removable by some monotonic data transformation.

In studies where an interaction effect has an important theoretical implication, a more compelling theoretical argument can be made if it can be shown, based on confidence intervals for the simple main effects, that the population interaction effect is non-removable. If all of the confidence intervals for the population simple main effects of Factor A exclude 0 and indicate that these population simple main effects have different signs, then the interaction effect is non-removable. Alternatively, if all of the confidence intervals for the population simple main effects of Factor B exclude 0 and indicate that these population simple main effects have different signs, then the interaction effect is non-removable.

3.27 Data Visualization

The results of a one-factor design can be illustrated using a bar chart or error bar chart with confidence interval lines as explained in section 2.16. The results of a two-factor design can be illustrated using a *clustered bar chart* where the means (or any other estimate) for the levels of one factor are represented by a cluster of bars (with different colors, shades, or patterns) and the levels of the second factor are represented by different clusters of bars. The spacing of the bars between clusters is larger than the spacing between the within-cluster bars. The range of values for the vertical axis should be selected in a way that does not give an exaggerated impression of the parameter estimates (see section 2.16).

In a clustered bar chart or error bar chart, it is easier to visually compare parameter estimates within clusters than across clusters. For example, in the clustered bar chart below it is easy to see the difference in estimated means between the two levels of Factor A within each level of Factor B. If the study has two treatment factors but one factor is more interesting than the other factor, the more interesting factor should be used as the within-cluster factor. If the study has one treatment factor and one classification factor, the treatment factor is usually the most interesting factor and would be used as the within-cluster factor. If two factors are equally interesting, use the factor with the largest eta-squared as the within-cluster factor. Alternatively, if two factors are equally interesting, a clustered bar chart or error bar chart will be more compact and easier to interpret if the factor with the largest number of levels is used as the within-cluster factor. If the A × B interaction effect is small, a bar chart for the main effects of Factor A and a second bar chart for the main effects of Factor B will convey the most important information more effectively than a clustered bar chart. The guidelines for using the confidence interval lines in a bar chart (section 2.16) to visually assess a hypothesis also apply to clustered bar charts.

A clustered bar chart and a clustered error bar chart for a 2 × 2 design with 95% confidence intervals are shown below using the data from Data Analysis Problem 3.3.

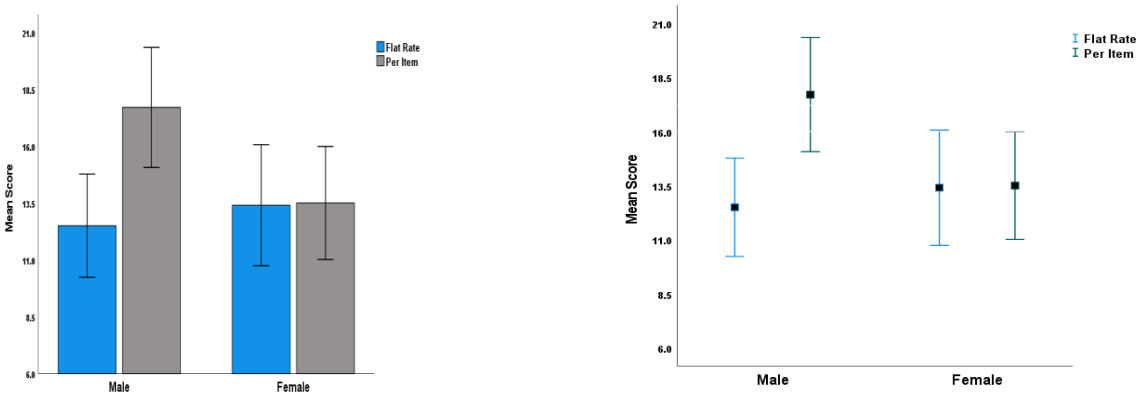


Figure 3.1 Clustered bar chart and alternative clustered error bar chart

Dot plots and a side-by-side box plot can also be used in one-factor and factorial designs to assist in the data screening process. A dot plot can be examined for the *y* scores within each level of a one-factor design and each combination of levels in a factorial design. In a factorial design, each factor combination is treated as a separate group in a side-by-side box plot.

Key Terms

between-subjects factor
treatment factor
classification factor
pairwise comparisons
Tukey-Kramer method
Games-Howell method
family-wise directional error rate (FWDER)
linear contrast
contrast coefficient
standardized linear contrast
Holm test
one-way ANOVA
F test
Fisher protected test
mean squared error
eta-squared

adjusted eta-squared estimate
2 × 2 factorial experiment
main effect
interaction effect
simple main effect
pairwise main effect comparisons
pairwise interaction effects
pairwise simple main effects
two-way ANOVA
partial eta-squared
generalized eta-squared
three-factor experiment
three-way interaction
simple-simple main effect
simple two-way interaction
three-way ANOVA
random factor
one-way random effects ANOVA
two-stage cluster sampling
Welch F test
Kruskal-Wallis test
removable interaction
non-removable interval
clustered bar chart

Concept Questions

1. What are the advantages of computing confidence intervals for all pairwise differences among population means instead of reporting the p-value for a one-way ANOVA?

2. What are the assumptions for the F test of equal population means in the one-way ANOVA?

3. What are the assumptions of an equal-variance confidence interval for a linear contrast of population means?

4. What are the assumptions of an unequal-variance confidence interval for a linear contrast of population means?

5. What are the assumptions for a confidence interval for eta-squared?

6. If two-group design is analyzed using a one-way ANOVA, how will the *F*-ratio and its *p*-value compare with the equal-variance *t*-statistic and its *p*-value?

7. Explain how the error variance can be reduced by including a classification factor.

8. What statistical method can be used to test H_0: $\mu_1 = \mu_2 = \mu_3$ in a between-subjects design?

9. If the AB interaction effect in a two-factor design is large, what effects would you want to estimate and what effects may not be interesting?

10. If the test for the AB interaction effect in a two-factor design is inconclusive, what effects would you want to estimate and what effects may not be interesting?

11. How can you show in a convincing way that four population means are approximately equal? Explain why a "nonsignificant" one-way ANOVA result does not provide evidence of similarity.

12. Draw a clustered bar chart of population means in a 2 × 2 design where the interaction effect is large. Draw a second bar chart where the interaction effect is zero but one of the main effects is large.

(For questions 13-17) Consider the following 2 × 2 factorial design with population means given in each cell.

		Factor B	
		b_1	b_2
Factor A	a_1	μ_1	μ_2
	a_2	μ_3	μ_4

13. Define the main effect of A in terms of the population means and specify the contrast coefficients that would define this effect.

14. Define the main effect of B in terms of the population means and specify the contrast coefficients that would define this effect.

15. Define the AB interaction effect in terms of the population means and specify the contrast coefficients that would define this effect.

16. Define the simple main effect of A at b_1 in terms of the population means and specify the contrast coefficients that would define this effect.

17. Define the simple main effect of B at a_2 in terms of the population means and specify the contrast coefficients that would define this effect.

18. Consider a one-way ANOVA summary table for a 4-group experiment with ten participants per group. Fill in the missing values.

Source	*SS*	*df*	*MS*	*F*
A	210	___	___	___
ERROR	___	___	___	
TOTAL	570			

19. Suppose a researcher uses a one-way ANOVA to test H_0: $\mu_1 = \mu_2 = \mu_3$ and reports that the results are "significant". What does this mean?

20. For a 2 × 2 factorial design, give examples of the four population means that would produce the following conditions:

a) large main effect of A, zero main effect of B, zero AB interaction
b) zero main effect of A, zero main effect of B, zero AB interaction
c) large main effect of A, large main effect of B, zero AB interaction
d) zero main effect of A, zero main effect of B, large AB interaction
e) large main effect of A, zero main effect of B, large AB interaction

21. Consider a two-way ANOVA summary table for a 4 × 5 factorial experiment with 5 participants per group. Fill in the missing values.

Source	*SS*	*df*	*MS*	*F*
A	15	___	___	___
B	20	___	___	___
AB	24	___	___	___
ERROR	___	___	___	
TOTAL	159			

22. If the test for an ABC interaction effect is inconclusive, what effects would you want to examine next?

23. If an ABC interaction effect is detected, what effects would you want to examine next?

24. If the tests for the ABC, AB, AC, and BC interaction effects are all inconclusive, what effects would you want to examine?

25. If an AB interaction effect is detected but the tests for the ABC, AC and BC interactions are inconclusive, what effects would you want to examine next?

26. When would a test or confidence interval for a linear contrast of population medians be preferred to a test or linear contrast of population means?

27. What is the distribution-free alternative to the one-way ANOVA *F* test?

28. What is an unequal-variance alternative to the one-way ANOVA *F* test?

29. Explain how you could show that an A × B interaction effect in a 2 × 2 design is non-removable.

30. What are the advantages and disadvantages of using a random factor rather than a fixed factor?

Data Analysis Problems

3-1. Twenty-one participants were randomly selected from a university research participant pool of about 4,500 students. The 21 participants were randomly assigned to three groups of equal size. All participants viewed the same very short (2-second) video of a moving car hitting a stopped car and were later asked to estimate the speed (in MPH) of the moving car when the moving car "*bumped* into the stopped car" (group 1), "*crashed* into the stopped car" (group 2) or "*smashed* into the stopped car" (group 3). The MPH estimates are shown below.

Bumped:	17	8	10	12	16	14	13
Crashed:	12	22	20	12	19	18	20
Smashed:	25	26	22	32	30	22	27

a) Describe the study population.

b) Describe the population means (μ_1, μ_2, and μ_3) in the context of this study.

c) Use SPSS or R to test H_0: $\mu_1 = \mu_2 = \mu_3$ with $\alpha = .05$. Report *F*, degrees of freedom, and *p*-value in APA style.

d) Compute unequal variance Tukey-Kramer (using R) or Games-Howell (using SPSS) pairwise confidence intervals and interpret the results.

e) Use SPSS or R to compute a 95% confidence interval for $\mu_3 - (\mu_1 + \mu_2)/2$ and interpret this result. Do not assume equal population variances.

f) Use SPSS or R to compute a 95% confidence interval for the standardized contrast $\mu_3 - (\mu_1 + \mu_2)/2$ and interpret this result.

g) Use SPSS or R to compute a 95% confidence interval for the population eta-squared (η^2) and interpret this result.

h) Use SPSS to test H_0: $\mu_1 = \mu_2 = \mu_3$ with $\alpha = .05$ using the Welch F test. Report F, degrees of freedom, and p-value in APA style.

i) How many participants would be needed in a future replication of this study to obtain Bonferroni 95% confidence intervals for all pairwise differences in population means with widths of 5.0? Use the MS_E from this study as the planning value of the average within-group error variance.

3-2. It has been suggested that men are more motivated by money than women. Twenty male students and twenty female students were randomly selected from a university research participant pool consisting of about 2,100 men and 2,200 women. The sample of male participants and the sample of female participants were each randomized into two groups of equal size. In the first group, participants were paid \$5 to show up for the study and an additional \$25 to complete a 30-item SAT-type math test. In the second group, participants were paid \$5 to show up and an additional \$2 for every question they answered correctly. The test scores for each of the four groups are shown below.

Factor A: Gender	Factor B: Payment	Test Score
Male	Flat rate	**14 15 11 7 16 12 15 16 10 9**
	Per item	**18 24 14 18 22 21 16 17 14 13**
Female	Flat rate	**16 11 10 17 13 18 12 16 6 15**
	Per item	**18 17 11 9 9 13 18 15 14 11**

a) Describe the two study populations.

b) Describe the population means (μ_{11}, μ_{12}, μ_{21}, and μ_{22}) in the context of this study.

c) Use SPSS or R to test the interaction effect in a two-way ANOVA. Report *F*, degrees of freedom, and *p*-value for the interaction effect in APA style. What effects should be examined next?

d) Use SPSS or R to compute Bonferroni 95% confidence intervals for the simple main effects of Payment at Male and at Female. Interpret the results. Do not assume equal population variances.

e) Use SPSS or R to produce a clustered bar chart with Payment as the within-cluster factor. Include 95% confidence interval bars.

f) How many participants would be needed in a future study to conduct Bonferroni tests of the two simple main effects of Payment with power of .95, α = .05, and an effect size of 3.0? Use the MS_E from this study as a planning value of the average within-group error variance.

3-3. Electronic versions of textbooks do not need to follow the same printing conventions as hard-copy textbooks. A traditional textbook is printed using techniques that reduce the number of pages and printing costs. It has been suggested that electronic textbooks could present text in nontraditional ways that might increase reading speed and comprehension. A 2 × 2 × 2 factorial experiment was conducted with text presented on a computer screen using an Arial or Times font style in a 12-point or 10-point font size with left justification or full justification. Forty students were randomly selected from a university research participant pool of about 5,000 students and were randomly assigned into eight groups of equal size. Each participant was presented with a short story in one of the eight possible text formats. The time (in seconds) to read the story was recorded for each participant. The reading times are given below.

Style	Size	Justification	Reading Time (seconds)				
Arial	12	Left	**305**	**300**	**305**	**270**	**240**
		Full	**259**	**292**	**298**	**297**	**293**
Arial	10	Left	**245**	**251**	**210**	**214**	**290**
		Full	**255**	**280**	**235**	**295**	**293**
Times	12	Left	**324**	**292**	**258**	**301**	**298**
		Full	**260**	**337**	**300**	**344**	**345**
Times	10	Left	**240**	**264**	**255**	**285**	**290**
		Full	**300**	**300**	**259**	**325**	**336**

a) Describe the study population

b) Use SPSS or R to test all main effects and interaction effects in a three-way ANOVA. For which effects can the null hypothesis be rejected and what effects have inconclusive results?

c) Use SPSS or R to compute Bonferroni 95% confidence intervals for the three main effects and interpret the results. Do not assume equal variances.

d) The researcher wants to replicate this study using a larger sample size. About how many participants are needed in each group to obtain Bonferroni 95% confidence intervals with widths of 20 seconds for the three main effects? Use the MS_E from this study as a planning value of the average within-group error variance.

Suggested Readings

Howell, D. C. (2013). *Statistical methods for psychology (8th ed.)* Thompson & Wadsworth. **Chapters 11, 12, and 13**

Kepple, G. & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook (4th ed).* Prentice Hall. **Parts II and III**

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2024). *Designing and analyzing data. A model comparison perspective (3rd ed)*. Lawrence Erlbaum. **Chapters 3 - 5, 7, and 8**

Snedecor, G. W. & Cochran, W. G. (1989). *Statistical methods (8th ed.).* Iowa State Press. **Chapters 12, 14, and 16**

Chapter 4

Within-subject Designs

4.1 Within-subject Experiments

An experiment where each participant is measured under all $a \geq 2$ treatment conditions is called a *within-subjects experiment*. A within-subjects experiment with $a = 2$ levels is also called a *paired-samples design*. The independent variable in a within-subjects experiment is called a *within-subjects factor*. With the same participants used in all treatment conditions, the total sample size in a within-subjects experiment to achieve desired power or precision can be substantially smaller than a comparable between-subjects experiment where participants are randomly divided into different groups and each participant receives only one treatment.

Another type of within-subjects experiment uses pairs of similar participants and randomly assigns one member of the pair to Treatment 1 and the other member to Treatment 2. Some examples of naturally occurring pairs are twins, married couples, and roommates. Alternatively, a sample of n participants (where n is an even number) could be arranged into $n/2$ *matched pairs* of participants who have similar quantitative characteristics. For example, the two youngest participants are matched to form pair 1, the next two youngest participants are matched to form pair 2, and so on.

In a 2-level within-subjects experiment, the goal is to estimate $\mu_1 - \mu_2$ where μ_1 is the mean of the response variable if everyone in the study population had received Treatment 1, and μ_2 is the mean of the response variable if everyone in the study population had received Treatment 2. These interpretations of μ_1 and μ_2 assume no *practice effect*, *fatigue effect*, or *carryover effect*. A practice effect will make the second treatment appear better than it is, and a fatigue effect will make the second treatment appear worse than it is. A carryover effect occurs when the effect of one treatment persists during exposure to a second treatment. When carryover occurs, μ_2 will reflect the combined effects of Treatment 1 and Treatment 2. If there are no

practice, fatigue, or carryover effects, the interpretations of μ_1 and μ_2 in a within-subjects experiment are the same as in a between-subjects experiment. If practice, fatigue, or carryover effects are possible, then treatment counterbalancing (described in Section 4.18) should be used.

4.2 Confidence Interval for a Population Mean Difference

Consider a random sample of n participants who have been measured under two treatment conditions. The two measurements for participant i are y_{i1} and y_{i2}. Compute a difference score $d_i = y_{i1} - y_{i2}$ for each of the n participants. Let $\hat{\mu}_d$ be the sample mean of the n difference scores and let $\hat{\sigma}_d^2$ be the sample variance of the n difference scores. It can be shown that $\hat{\mu}_d = \hat{\mu}_1 - \hat{\mu}_2$ (i.e., the mean of the difference scores is equal to a difference in the means). A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\hat{\mu}_d \pm t_{\alpha/2;df}\sqrt{\frac{\hat{\sigma}_d^2}{n}} \tag{4.1}$$

where $df = n - 1$ and $\sqrt{\hat{\sigma}_d^2/n}$ is the estimated standard error of $\hat{\mu}_1 - \hat{\mu}_2$. SPSS and R will compute Equation 4.1. Note that Equation 4.1 is the same as Equation 1.6 applied to difference scores. The interpretation of a confidence interval for $\mu_1 - \mu_2$ in a within-subjects experimental design and a two-group experimental design are the same if there are no practice, fatigue, or carryover effects.

It can be shown (see Appendix B.5 of Volume 2) that $\hat{\sigma}_d^2 = \hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\hat{\rho}_{12}\hat{\sigma}_1\hat{\sigma}_2$ where $\hat{\rho}_{12}$ is the sample Pearson correlation between the two measurements (a Pearson correlation is a measure of association between two quantitative variables that has a range of -1 to 1). From this equation we see that the variance of the difference scores is smaller for larger values of $\hat{\rho}_{12}$. It is common for the two measurements in a within-subjects experiment to be moderately or highly correlated and $\hat{\sigma}_d^2$ is often much smaller than the variance of either y_1 or y_2.

4.3 Confidence Interval for a Population Standardized Mean Difference

The population standardized mean difference in a within-subjects experiment is defined in exactly the same way as in a between-subjects experiment (see Equation

2.2). A standardized mean difference may be easier to explain than $\mu_1 - \mu_2$ in applications where the scale of the response variable is not familiar to the intended audience.

A $100(1-\alpha)\%$ confidence interval for $\delta = (\mu_1 - \mu_2)/\sqrt{(\sigma_1^2 + \sigma_2^2)/2}$ in a paired-samples design is

$$\hat{\delta} \pm z_{\alpha/2} SE_{\hat{\delta}} \tag{4.2}$$

where $\hat{\delta} = (\hat{\mu}_1 - \hat{\mu}_2)/\hat{\sigma}$, $SE_{\hat{\delta}} = \sqrt{\frac{\hat{\delta}^2(\hat{\sigma}_1^4 + \hat{\sigma}_2^4 + 2\hat{\rho}_{12}^2\hat{\sigma}_1^2\hat{\sigma}_2^2)}{8(n-1)\hat{\sigma}^4} + \frac{\hat{\sigma}_d^2}{(n-1)\hat{\sigma}^2}}$, $\hat{\sigma} = \sqrt{(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2}$, and $\hat{\sigma}_d^2 = \hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\hat{\rho}_{12}\hat{\sigma}_1\hat{\sigma}_2$. Equation 4.2 does not assume equal population variances. An alternative to δ is $\delta'' = (\mu_1 - \mu_2)/\sigma_j$ (j = 1 or 2) and is easier to interpret than δ when the population variances are not similar. The `ci.stdmean.ps` function in the `statpsych` package will compute a confidence interval for δ and δ''.

Example 4.1. Eight participants were randomly selected from a study population of about 1,500 college students. Participants pressed button 1 as soon as they saw the letter E displayed on a computer screen and pressed button 2 if the letter E was displayed backwards. The Es and backwards Es were displayed two ways: upright (zero rotation) and rotated 60^0. The reaction times (in milliseconds) to the backwards Es under the upright and rotated conditions are given below along with the difference scores.

									$\hat{\mu}$	$\hat{\sigma}$
0^0:	621	589	604	543	588	647	638	589	602.4	33.17
60^0:	704	690	741	635	724	736	780	635	705.6	51.08
d:	-83	-101	-137	-92	-136	-89	-142	-46	$\hat{\rho}_{12}$ = .769	

The 95% confidence interval for $\mu_1 - \mu_2$ is [-131, -75]. The researcher is 95% confident that the mean reaction time to the rotated backwards E is 75 to 131 ms longer than the mean reaction time to the non-rotated backwards E in the study population of 1,500 college students. The 95% confidence interval for δ is [-3.67, -1.12]. The researcher is 95% confident that the mean reaction time to the rotated backwards E is 1.12 to 3.67 standard deviations greater than the mean reaction time to the non-rotated backwards E in the study population of 1,500 college students. In this example, the researcher would report the confidence interval for $\mu_1 - \mu_2$ rather than the confidence interval for δ because reaction time in milliseconds is an easily understood scale.

The estimate of δ has a slight positive bias in small samples. The bias can be reduced by multiplying $\hat{\delta}$ by $\sqrt{(n-2)/(n-1)}$. The bias adjustment for δ'' is $1 - 3/[4n-5]$. A bias adjusted estimate is typically reported along with a

confidence interval, but the bias adjusted estimate is not needed or recommended in the computation of the confidence interval formula.

4.4 Confidence Interval for a Ratio of Population Means

If the response variable is measured on a ratio scale, a ratio of population means μ_1/μ_2 is a unitless measure of effect size that could be more meaningful and easier to interpret than a standardized mean difference. An approximate $100(1-\alpha)\%$ confidence interval for μ_1/μ_2 is

$$exp[ln(\hat{\mu}_1/\hat{\mu}_2) \pm t_{\alpha/2;df}\sqrt{\frac{\hat{\sigma}_1^2}{\hat{\mu}_1^2 n} + \frac{\hat{\sigma}_2^2}{\hat{\mu}_2^2 n} - \frac{2\hat{\rho}_{12}\hat{\sigma}_1\hat{\sigma}_2}{\hat{\mu}_1\hat{\mu}_2 n}}\,] \tag{4.3}$$

where $df = n - 1$. The `ci.ratio.mean.ps` function in the `statpsych` package will compute Equation 4.3.

Example 4.2. A sample of 12 college students were shown a cartoon face on the left side of a computer screen and were asked to draw a copy of the face on the right side of the screen. Then the students were shown the same face upside down and were asked to draw a copy of the inverted face. A drawing error score for each participant was computed from angular differences between points on the drawn face and corresponding points on the target face. These error scores do not have a simple interpretation and a ratio of means serves as a useful unitless measure of effect size. The sample data are given below

Inverted: `76.41 66.91 81.06 74.78 83.76 89.31 78.78 87.06 82.61 76.74 88.33 86.18`
Upright: `59.85 60.64 84.86 68.16 71.53 86.18 67.30 65.46 83.50 66.76 88.37 65.02`

A 95% confidence interval for μ_1/μ_2 is [1.04, 1.21]. This result indicates that the population mean drawing error for the inverted face is 1.04 to 1.21 times as large as the population mean drawing error for the upright face.

4.5 Prediction Interval for a Score Difference

In some paired-samples experiments, the researcher might want to estimate how the response variable score for *one* randomly selected person would differ under the two treatment conditions. A $100(1-\alpha)\%$ prediction interval for this difference in a paired-samples experiment is

$$\hat{\mu}_1 - \hat{\mu}_2 \pm t_{\alpha/2;df}\sqrt{\hat{\sigma}_d^2 + \hat{\sigma}_d^2/n} \tag{4.4}$$

where $df = n - 1$ and $\hat{\sigma}_d^2 = \hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\hat{\rho}_{12}\hat{\sigma}_1\hat{\sigma}_2$. A prediction interval for the difference in scores for one person will never be narrower, and is often much wider

than a confidence interval for the difference in population means. The `pi.score` function in the `statpsych` package can be used to compute Equation 4.4 by first computing the mean and standard deviation of the difference scores.

Example 4.3. In *Example 4.1*, the 95% confidence interval for $\mu_1 - \mu_2$ is [-131, -70]. A 95% prediction interval for a difference in reaction time scores for a single student in this paired-samples experiment is [-186.6, -19.9]. This prediction interval suggests that the reaction time for a randomly selected student would be 19.9 to 186.6 ms longer to a rotated backwards E than a non-rotated backwards E.

4.6 Linear Contrasts

In a within-subjects study with $a \geq 2$ levels, participant i produces $a \geq 2$ scores ($y_{i1}, y_{i2}, \ldots, y_{ia}$) and a *linear contrast score* for participant i is

$$l_i = \sum_{j=1}^{a} q_j y_{ij} \tag{4.5}$$

where q_j are contrast coefficients specified by the researcher. Note that l_i specializes to a difference score (d_i) when one coefficient is 1, another coefficient is -1, and all other coefficients are zero. It can be shown that the mean of the linear contrast scores ($\hat{\mu}_l$) is equal to a linear contrast of sample means ($\sum_{j=1}^{a} q_j \hat{\mu}_j$)

$$\hat{\mu}_l = \sum_{j=1}^{a} q_j \hat{\mu}_j \tag{4.6}$$

and the estimated variance of the linear contrast scores is

$$\hat{\sigma}_l^2 = \sum_{i=1}^{n} (l_i - \hat{\mu}_l)^2/(n-1). \tag{4.7}$$

A $100(1-\alpha)\%$ confidence interval for $\sum_{j=1}^{a} q_j \mu_j$ is

$$\hat{\mu}_l \pm t_{\alpha/2;df} \sqrt{\frac{\hat{\sigma}_l^2}{n}} \tag{4.8}$$

where $df = n - 1$ and $\sqrt{\hat{\sigma}_l^2/n}$ is the estimated standard error of $\sum_{j=1}^{a} q_j \hat{\mu}_j$. A Bonferroni adjustment to α in the critical t-value of Equation 4.8 can be used when two or more simultaneous confidence intervals are required. Note that Equation 4.8 is Equation 1.6 applied to linear contrast scores.

Example 4.4. Six participants were randomly selected from a study population of about 870 4th year psychology majors and were given detailed descriptions of the human resource departments at four San Francisco based manufacturing companies. The first two companies produced electronic products, and the second two companies produced clothing products. The students were asked to rate their interest in working in the human resources department of each company on a 1 to 60 scale. The ratings are given below.

Student	Company 1	Company 2	Company 3	Company 4	l
1	24	27	21	20	5.0
2	37	35	31	29	6.0
3	20	18	17	18	1.5
4	45	48	40	41	6.0
5	49	52	43	40	9.0
6	32	34	30	27	4.5

The researcher wants a 95% confidence interval for the linear contrast $(\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4)/2$. The linear contrast score $l_i = (y_{i1} + y_{i2})/2 - (y_{i3} + y_{i4})/2$ was computed and is reported in the last column. A 95% confidence interval for $(\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4)/2$ is [2.8, 7.9]. The researcher is 95% confident that the mean rating, averaged over the two electronic manufacturing companies, is 2.8 to 7.9 points greater than the mean rating, averaged over the two clothing manufacturing companies, in the study population of 870 4th year college students.

4.7 Standardized Linear Contrasts

In applications where a linear contrast of means could be difficult to interpret because the scale of the response variable is not familiar to the intended audience, it may be helpful to report a confidence interval for a standardized linear contrast of population means

$$\varphi = \sum_{j=1}^{a} q_j \mu_j / \sqrt{\sum_{j=1}^{a} \sigma_j^2 / a} \tag{4.9}$$

which is a generalization of Equation 4.2. A $100(1-\alpha)\%$ confidence interval for φ that assumes equal population variances and equal population correlations is

$$\hat{\varphi} \pm z_{\alpha/2} SE_{\hat{\varphi}} \tag{4.10}$$

where $\hat{\varphi} = \sum_{j=1}^{a} q_j \hat{\mu}_j / \sqrt{\sum_{j=1}^{a} \hat{\sigma}_j^2 / a}$, $SE_{\hat{\varphi}} = \sqrt{\frac{\hat{\varphi}^2[1 + (a-1)\hat{\rho}^2]}{2a(n-1)} + \frac{(1-\hat{\rho})\sum_{j=1}^{a} q_j^2}{n-1}}$, and $\hat{\rho}$ is the average of the sample correlations for the $a(a-1)/2$ pairs of measurements. The `ci.lc.stdmean.ws` function in the `statpsych` package will compute Equation 4.10 using an alternative standard error that assumes equal population correlations but does not assume equal population variances.

4.8 Directional Two-sided Test

The confidence interval for $\mu_1 - \mu_2$ (Equation 4.1) can be used to perform a directional two-sided test of the following hypotheses.

H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 > \mu_2$ H_2: $\mu_1 < \mu_2$

If the lower confidence limit for $\mu_1 - \mu_2$ is greater than 0, then reject H_0 and accept H_1: $\mu_1 > \mu_2$; if the upper confidence limit for $\mu_1 - \mu_2$ is less than 0, then reject H_0 and accept H_2: $\mu_1 < \mu_2$. The results are inconclusive if the confidence interval includes 0.

A directional two-sided test for a difference in population means in a paired-samples design can also be performed using a *paired-samples t-test*. SPSS and R will compute the test statistic

$$t = (\hat{\mu}_1 - \hat{\mu}_2)/SE_{\hat{\mu}_1 - \hat{\mu}_2} = \hat{\mu}_d/\sqrt{\hat{\sigma}_d^2/n} \tag{4.11}$$

and its associated *p*-value. If the *p*-value is less than α, then H_0 is rejected and then either H_1 or H_2 is accepted depending on the sign of *t*. Remember that a *p*-value less than .05 does not imply that there is an important difference in the two population means, and a *p*-value greater than .05 does not imply that the two population means are equal.

The confidence interval for $\sum_{j=1}^{a} q_j\mu_j$ (Equation 4.8) can be used to perform a directional two-sided test of the following hypotheses.

H_0: $\sum_{j=1}^{a} q_j\mu_j = 0$ H_1: $\sum_{j=1}^{a} q_j\mu_j > 0$ H_2: $\sum_{j=1}^{a} q_j\mu_j < 0$

If the lower confidence limit for $\sum_{j=1}^{a} q_j\mu_j$ is greater than 0, reject H_0 and accept H_1: $\sum_{j=1}^{a} q_j\mu_j > 0$; if the upper confidence limit for $\sum_{j=1}^{a} q_j\mu_j$ is less than 0, reject H_0 and accept H_2: $\sum_{j=1}^{a} q_j\mu_j < 0$. The results are inconclusive if the confidence interval includes 0.

The test statistic for the one-sample *t*-test described in Chapter 1, when applied to linear contrast scores, is $t = (\hat{\mu}_l - 0)/\sqrt{\hat{\sigma}_l^2/n}$ and is equal to $t = \sum_{j=1}^{a} q_j\hat{\mu}_j/SE_{\sum_{j=1}^{a} q_j\hat{\mu}_j}$. The *p*-value for the one-sample *t*-test is used to decide if H_0: $\sum_{j=1}^{a} q_j\mu_j = 0$ can be

rejected or not. If H_0 can be rejected, then H_1 or H_2 is accepted depending on the sign of t. For a directional two-sided test of H_0: $\mu_1 = \mu_2$ or H_0: $\sum_{j=1}^{a} q_j \mu_j = 0$, the confidence interval approach and the test statistic approach always give the same conclusion.

The probability of making a directional error using a single paired-samples t-test or a single test of a linear contrast is at most $\alpha/2$. If two or more hypotheses are tested, the Holm method described in Chapter 3 can be used to keep the FWDER below $\alpha/2$.

4.9 Equivalence Test

A confidence interval for $\mu_1 - \mu_2$ in a paired-samples design can be used to choose between the following two hypotheses in an equivalence test

$$H_1: |\mu_1 - \mu_2| < h \qquad\qquad H_2: |\mu_1 - \mu_2| \geq h$$

where h is some value specified by the researcher. Usually, h represents the value of $\mu_1 - \mu_2$ that would be considered by experts to be small or unimportant. As in section 2.8, the interval $-h$ to h defines a region of practical equivalence. If the confidence interval for $\mu_1 - \mu_2$ is completely contained within the range $-h$ to h, then H_1 is accepted; if the confidence interval for $\mu_1 - \mu_2$ is completely outside the range $-h$ to h, then H_2 is accepted; otherwise, the results are inconclusive. The probability of falsely accepting H_1: $|\mu_1 - \mu_2| < h$ is at most $\alpha/2$. Equivalence tests typically use $\alpha = .10$ rather than $\alpha = .05$.

Example 4.5. Fifty participants were randomly selected from a study population of about 12,500 college students and were then shown two different 30-second public service announcements about HIV pre-exposure prophylaxis. The two announcements used different actors and settings. After viewing both announcements, each participant rated the effectiveness of each announcement on a 0 to 10 scale. If the population mean difference is less than 1, both announcements will be used. The 95% confidence interval for the population mean difference is [-0.87, 0.650] which suggests that both announcements have similar perceived effectiveness and both will be used.

4.10 Superiority and Noninferiority Tests

The superiority and noninferiority tests described in section 2.9 for two-group experiments can also be applied to paired-samples experiments. If the goal of the study is to show that some treatment is meaningfully better than another treatment, then the hypotheses can be set up in the form of a superiority test. If the goal of the study is to show that a less costly treatment is not meaningfully worse than a more costly treatment, then the hypotheses can be set up in the form of a noninferiority test. The confidence interval for $\mu_1 - \mu_2$ given in Equation 4.1 can be used to perform a superiority or a noninferiority test. As in the case of a 2-group design, the probability of making a directional error in a noninferiority test or a superiority test in a paired-samples design is at most $\alpha/2$.

In a test of noninferiority or superiority where it is difficult to specify a value of $\mu_1 - \mu_2$ that would be considered small or unimportant, it might be easier to specify a region of practical equivalence for δ. A confidence interval for δ using Equation 4.2 could then be used to decide of Treatment 1 is noninferior to Treatment 2 in a noninferiority test or decide if Treatment 1 is superior to Treatment 2 in a superiority test.

4.11 One-way Within-subjects Analysis of Variance

The variability of the y scores across participants and treatments in a one-factor within-subjects study can be decomposed into three sources of variability as shown in the within-subjects ANOVA table below. The levels of "Factor S" represent the n participants. Factor S is a random factor because the n participants are assumed to be a random sample from some study population of potential participants. Note that the *MS* for the AS interaction is used as the error variance for the main effect of Factor A.

Source	*SS*	*df*	*MS*	*F*
A	SS_A	$df_A = a - 1$	$MS_A = SS_A/df_A$	MS_A/MS_{AS}
S	SS_S	$df_S = n - 1$	$MS_S = SS_S/df_S$	
AS	SS_{AS}	$df_{AS} = (a - 1)(n - 1)$	$MS_{AS} = SS_{AS}/df_{AS}$	

The sum of squares in the ANOVA table can be computed using the following formulas

$$SS_A = n\sum_{j=1}^{a}\left(\hat{\mu}_{+j} - \hat{\mu}_{++}\right)^2 \quad (4.12)$$

$$SS_S = a\sum_{i=1}^{n}(\hat{\mu}_{i+} - \hat{\mu}_{++})^2 \quad (4.13)$$

$$SS_{AS} = \sum_{j=1}^{a}\sum_{i=1}^{n}\left(y_{ij} - \hat{\mu}_{++}\right)^2 - SS_A - SS_S \quad (4.14)$$

where $\hat{\mu}_{++} = \sum_{j=1}^{a}\sum_{i=1}^{n} y_{ij}/an$, $\hat{\mu}_{+j} = \sum_{i=1}^{n} y_{ij}/n$ and $\hat{\mu}_{i+} = \sum_{j=1}^{a} y_{ij}/a$.

The recommended estimate of η_A^2 for the within-subjects factor is

$$\hat{\eta}_A^2 = SS_A/(SS_A + SS_S + SS_{AS}). \quad (4.15)$$

Note that Equation 4.15 is a generalized eta-squared estimate and not a partial eta-squared estimate. Equation 3.19 can be used to reduce the positive bias of $\hat{\eta}_A^2$.

The *F* statistic from the ANOVA table can be used to test the omnibus null hypothesis H_0: $\mu_1 = \mu_2 = \ldots = \mu_a$. The alternative hypothesis states that there is at least one pair of population means that are not identical. SPSS and R will report a *p*-value for the *F* statistic that is used to decide if the null hypothesis can be rejected. A test of this hypothesis in the within-subjects ANOVA suffers from the same problem as the test of equal population means in the between-subjects ANOVA. Specially, we know with near certainty that H_0: $\mu_1 = \mu_2 = \ldots = \mu_a$ will almost never be true (or equivalently, we know that η_A^2 will almost never be exactly zero). Thus, a statistical test that simply rejects or fails to reject the null hypothesis of equal population means does not provide useful scientific information. Pairwise comparisons or other linear contrasts of the within-subjects means (standardized or unstandardized) are the recommended supplements or replacements to the within-subjects ANOVA omnibus test.

The *F* test is useful in the special case of $a = 3$ for conducting a Fisher protected test, which can be more powerful than the Holm test. If H_0: $\mu_1 = \mu_2 = \mu_3$ is rejected, then follow-up hypothesis tests of pairwise comparisons or linear contrasts can be performed without any adjustments to α and the FWDER will stay below $\alpha/2$.

4.12 Wide and Long Data Formats

Most confidence intervals and test statistics for within-subjects designs require the data to be in "wide format" but some analyses require the data to be in "long format". For example, some statistical packages will compute a one-way within-subjects ANOVA using data in wide format, but then the ANOVA table will not include the source due to participants (Factor S). The sum of squares for Factor A, Factor S, and the AS interaction (SS_A, SS_S, and SS_{AS}) are all needed to compute the generalized eta-squared estimate (Equation 4.15). The mean squares for Factor S and the AS interaction (MS_S, and MS_{AS}) are needed to compute the reliability estimates described in Section 4.26 (Equations 4.18 and 4.19).

To illustrate the difference between wide and long data formats, consider a study with n = 4 participants who are each measured under two treatment conditions. Hypothetical data in wide format are shown below.

Participant	Treatment 1	Treatment 2
1	10	14
2	15	18
3	12	13
4	14	20

In a wide format, the SPSS data file would have four rows with one row per participant. The same data in long format are shown below. In long format, the SPSS data file would have 8 rows.

Participant	Treatment	Score
1	1	10
1	2	14
2	1	15
2	2	18
3	1	12
3	2	13
4	1	14
4	2	20

4.13 Pretest-posttest Designs

Another type of within-subjects design is the *pretest-posttest design* where the response variable is measured on one or more occasions prior to treatment and on one or more occasions following treatment. The pretest-posttest design is a more

ethical alternative to the two-group experiment with a control group in applications where the treatment is expected to be beneficial. In the simplest version of this design, each participant is measured once prior to treatment (the pretest) and once following treatment (the posttest). The two measurements can be used to construct a confidence interval for $\mu_1 - \mu_2$ where μ_1 is the population mean of the response variable prior to treatment and μ_2 is the population mean of the response variable after treatment.

If the confidence interval excludes 0, it is tempting to conclude that the treatment caused the mean of the response variable to change. Recall from Chapter 2 that causality require an absence of any confounding variable. In a pretest-posttest design, there are two types of confounding variables: *time-invariant* confounding variables, and *time-varying* confounding variable. A time-invariant confounding variable remains constant over the pretest and posttest periods (e.g., gender, ethnicity). A *time-varying* confounding variable does not remain constant over the pretest and posttest periods. The pretest-posttest design is called a *quasi-experimental design* because it controls for all time-invariant confounding variables that do not interact with the treatment. The pretest-posttest design does not control for time-varying confounding variables.

If more than one pretest or more than one posttest measurements can be obtained, it may be possible to rule out certain time-varying confounding variables. For example, suppose that two pretest measurements (y_1 and y_2) and two posttest measurement (y_3 and y_4) are obtained for each participant. Furthermore, suppose that the treatment (which is given between Time 2 and Time 3) is expected to have a long term effect on the response variable. In this study, the researcher expects the difference between μ_1 and μ_2 to be small, the difference between μ_2 and μ_3 to be meaningfully large, and the difference between μ_3 and μ_4 to be small. If confidence intervals for these three pairwise comparisons are consistent with the predictions, then this would provide compelling evidence of a causal effect of treatment. This causal claim would be difficult to challenge because a skeptic would need to propose the existence of a specific time-varying confounding variable that also remained constant from Time 1 to Time 2, changed from Time 2 to Time 3, and then remained constant from Time 3 to Time 4.

In a pretest-posttest design with multiple pretests and posttests, the researcher can propose very specific changes in population means that would be expected if there was a causal effect of treatment. If these predictions are confirmed in a set of Bonferroni confidence intervals, then it might be reasonable to cautiously conclude that the treatment has a causal effect on the response variable.

Example 4.6. Thirty participants were randomly selected from a study population of about 2,700 first-year college students and were given a study skills questionnaire (scored 0 to 100). The 30 students then participated in a study skills workshop. Eight weeks after the workshop ended, the 30 students were given the study skills questionnaire again. The 95% confidence interval for $\mu_1 - \mu_2$ is [-7.24, -3.68]. The researcher is 95% confident that the mean study skills score would increase 3.38 to 7.24 points if all 2,700 first-year students participated in the workshop.

4.14 Two-factor Within-subjects Experiments

In a two-factor within-subjects experiment, all participants are measured under all combinations of the two factors. In the simple case of a 2 × 2 factorial experiment, all participants are measured under four conditions (a_1b_1, a_1b_2, a_2b_1, a_2b_2) with scores for participant i denoted as y_{i11}, y_{i12}, y_{i21}, and y_{i22}. The population means under these four conditions are μ_{11}, μ_{12}, μ_{21}, and μ_{22}. The main effects and interaction effects that were previously defined for a 2 × 2 between-subjects design also apply to a 2 × 2 within-subjects design as shown below.

AB interaction effect: $(\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22})$

Main effect of Factor A: $(\mu_{11} + \mu_{12})/2 - (\mu_{21} + \mu_{22})/2$

Main effect of Factor B: $(\mu_{11} + \mu_{21})/2 - (\mu_{12} + \mu_{22})/2$

The main effects could be misleading if there is an AB interaction effect. Simple main effects should usually be examined if an AB interaction effect has been detected. The simple main effects that were previously defined for a 2 × 2 between-subjects design also apply to a 2 × 2 within-subjects design as shown below.

A at b_1: $\mu_{11} - \mu_{21}$ B at a_1: $\mu_{11} - \mu_{12}$

A at b_2: $\mu_{12} - \mu_{22}$ B at a_2: $\mu_{21} - \mu_{22}$

Confidence intervals for the above effects are obtained by computing the appropriate linear contrast score for all n participants and then applying Equation 4.7. For example, to obtain a confidence interval for the AB interaction effect, the linear contrast score for participant i would be $l_i = y_{i11} - y_{i12} - y_{i21} + y_{i22}$. Alternatively, the `ci.2x2.mean.ws` and `ci.2x2.stdmean.ws` functions in the `statpsych` package will compute confidence intervals for the AB interaction, the two main effects, and the four simple main effects in a 2 × 2 within-subjects design.

In two-factor within-subject designs where one or both factors have more than two levels, pairwise main effects, pairwise interaction effects, or pairwise simple main effects could be examined. These pairwise effects are defined in exactly the same way as they were in the case of a two-factor between-subjects design. Consider the following table of population means for a 3 × 3 within-subjects design.

		Factor B		
		b_1	b_2	b_3
Factor A	a_1	μ_{11}	μ_{12}	μ_{13}
	a_2	μ_{21}	μ_{22}	μ_{23}
	a_3	μ_{31}	μ_{32}	μ_{33}

A few examples of effects that could be estimated in this design and their corresponding linear contrast scores are given below. A confidence interval for any of these effects is obtained by computing the appropriate linear contrast score for all n participants and then applying Equation 4.7.

Example 4.7a. A main effect pairwise comparison for levels 1 and 2 of Factor A is $(\mu_{11} + \mu_{12} + \mu_{13})/3 - (\mu_{21} + \mu_{22} + \mu_{23})/3$, and the corresponding linear contrast score for participant i is $l_i = (y_{i11} + y_{i12} + y_{i13})/3 - (y_{i21} + y_{i22} + y_{i23})/3$.

Example 4.7b. The simple main effect pairwise comparison for levels 1 and 2 of Factor A at level 1 of Factor B is $\mu_{11} - \mu_{21}$, and the corresponding linear contrast score for participant i is $l_i = y_{i11} - y_{i21}$.

Example 4.7c. The interaction pairwise comparison for levels 1 and 2 of Factor A and levels 2 and 3 of Factor B is $\mu_{12} - \mu_{13} - \mu_{22} + \mu_{23}$, and the corresponding linear contrast score for participant i is $l_i = y_{i12} - y_{i13} - y_{i22} + y_{i23}$.

4.15 Two-way Within-subjects Analysis of Variance

The variability of the y scores across participants and the two within-subject factors can be decomposed into seven sources of variability, as shown in the two-way within-subjects ANOVA table below. Note that the error variances for A, B, and AB are MS_{AS}, MS_{BS}, and MS_{ABS}, respectively.

Source	SS	df	MS	F
A	SS_A	$df_A = a - 1$	$MS_A = SS_A/df_A$	MS_A/MS_{AS}
B	SS_B	$df_B = b - 1$	$MS_B = SS_B/df_B$	MS_B/MS_{BS}
AB	SS_{AB}	$df_{AB} = (a - 1)(b - 1)$	$MS_{AB} = SS_{AB}/df_{AB}$	MS_{AB}/MS_{ABS}
S	SS_S	$df_S = n - 1$	$MS_S = SS_S/df_S$	
AS	SS_{AS}	$df_{AS} = (a - 1)(n - 1)$	$MS_{AS} = SS_{AS}/df_{AS}$	
BS	SS_{BS}	$df_{BS} = (b - 1)(n - 1)$	$MS_{BS} = SS_{BS}/df_{BS}$	
ABS	SS_{ABS}	$df_{ABS} = (a - 1)(b - 1)(n - 1)$	$MS_{ABS} = SS_{ABS}/df_{ABS}$	

The following generalized eta-squared estimates are recommended for a two-way within-subjects design. Equation 3.19 can be used to reduce the positive bias of these estimates.

$$\hat{\eta}_A^2 = SS_A/(SS_A + SS_S + SS_{AS} + SS_{BS} + SS_{ABS}) \quad (4.16a)$$

$$\hat{\eta}_B^2 = SS_B/(SS_B + SS_S + SS_{AS} + SS_{BS} + SS_{ABS}) \quad (4.16b)$$

$$\hat{\eta}_{AB}^2 = SS_{AB}/(SS_{AB} + SS_S + SS_{AS} + SS_{BS} + SS_{ABS}) \quad (4.16c)$$

A two-way within-subjects ANOVA can be used to test three omnibus null hypotheses: a zero main effect of Factor A, a zero main effect of Factor B, and a zero AB interaction effect. Hypothesis testing in the two-way within-subjects ANOVA suffers from the same problems as the two-way between-subjects ANOVA. Remember that a p-value less than .05 for a particular effect does not imply that the effect is large or important, and a p-value greater than .05 does not imply that the effect is zero. Confidence intervals for linear contrasts of means, or standardized linear contrasts of means are the recommended alternatives or supplements to the F tests. The same analysis strategy for a two-factor between-subjects design that was described in section 3.13 also applies to the two-factor within-subjects design.

4.16 Two-factor Mixed Designs

A *mixed two-factor design* has one between-subjects factor and one within-subjects factor. This design is also called a *split plot design.* The mixed design often provides greater power and narrower confidence intervals than a two-factor between-subjects design. The mixed design may be preferred to a two-factor within-subjects design if there is a concern of carryover effects for one of the factors. A mixed design may also be preferred in studies where the levels of one factor are most conveniently applied to each participant and the levels of the other factor are most conveniently or appropriately applied to different groups of participants. The two-factor mixed design is useful in studies where participants may have difficulty responding to all $a \times b$ levels of two within-subject factors. For example, instead of measuring participants under all six treatment conditions of a 2×3 within-subjects design, participants could be randomly divided into two groups with one group receiving treatment a_1, the other group receiving treatment a_2, and all participants receiving just three within-subject treatments b_1, b_2, and b_3.

The within-subjects factor can be a treatment factor, where participants are measured under all treatment conditions, or a pretest-posttest factor where participants are measured on two or more occasions before and after exposure to a single treatment. The between-subjects factor can be a treatment factor or a classification factor. Adding a classification factor to a within-subjects experiment can increase the generalizability of results for the treatment factor. The levels of the classification factor could represent different types of participants or different testing conditions.

The levels of the classification factor could also represent different sets of stimuli for the within-subjects conditions. In some within-subjects experiments, each participant responds to multiple stimuli (e.g., words, pictures, stories, etc.). When these stimuli are selected from a larger set of possible stimuli, there is always a concern that the results of the statistical analysis may not generalize to other stimuli. In these situations, greater generalizability can be achieved by adding a between-subjects factor with factor levels that represent different samples of stimuli with participants randomly assigned to the different stimuli sets.

The 2 × 2 mixed design is the simplest type of mixed two-factor design. Consider the following 2 × 2 mixed design where Factor A is the within-subjects factor and Factor B is the between-subjects factor.

		Factor B	
		b_1	b_2
Factor A	a_1	μ_{11}	μ_{12}
	a_2	μ_{21}	μ_{22}

All of the effects that were previously defined for the 2 × 2 between-subjects design and the 2 × 2 within-subjects design also apply to the 2 × 2 mixed design. Confidence intervals for these effects can be computed using a combination of methods and principles described previously. A confidence interval for the AB interaction is obtained by first computing a difference score for each participant (q_1 = 1 and q_2 = -1) and then applying Equation 2.1 (or Equation 3.1 with v_1 = 1 and v_2 = -1). The population mean of the difference scores for participants at level b_1 is $\mu_{11} - \mu_{21}$, and the population mean of the difference scores for participants at level b_2 is $\mu_{12} - \mu_{22}$. Thus, a confidence interval for the difference in population mean difference scores gives a confidence interval for $(\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22})$ which is the AB interaction effect. The analysis of the AB interaction effect in a 2 × 2 mixed design is also referred to as a "difference-in-differences" analysis.

A confidence interval for the main effect of Factor B (the between-subjects factor) is obtained by computing an average within-subjects score (q_1 = 1/2 and q_2 = 1/2) for each participant and then applying Equation 2.1 (or Equation 3.1 with v_1 = 1 and v_2 = -1). The population mean of the average of the two scores for participants at level b_1 is $(\mu_{11} + \mu_{21})/2$, and the mean of the average of the two scores for participants at level b_2 is $(\mu_{12} + \mu_{22})/2$. Thus, a confidence interval for the difference in population mean average scores gives a confidence interval for $(\mu_{11} + \mu_{21})/2 - (\mu_{12} + \mu_{22})/2$, which is the main effect of Factor B.

A confidence interval for the main effect of Factor A (the within-subjects factor) is obtained by computing a difference score for each participant (q_1 = 1 and q_2 = -1) and then applying Equation 3.1 with v_1 = 1/2 and v_2 = 1/2. The population mean of the difference scores at level b_1 is $\mu_{11} - \mu_{21}$, and the population mean of the

difference scores at level b_2 is $\mu_{12} - \mu_{22}$. Thus, a confidence interval for the average of these two population differences gives a confidence interval for $(\mu_{11} - \mu_{21})/2 + (\mu_{12} - \mu_{22})/2$ which is equal to $(\mu_{11} + \mu_{12})/2 - (\mu_{21} + \mu_{22})/2$.

A confidence interval for the simple main effect of A at b_1 is obtained by computing a difference score for each participant ($q_1 = 1$ and $q_2 = -1$) and applying Equation 1.6 to the participants at level b_1 (or use Equation 3.1 with $v_1 = 1$ and $v_2 = 0$). Likewise, a confidence interval for the simple main effect of A at b_2 is obtained by computing a difference score for each participant ($q_1 = 1$ and $q_2 = -1$) and applying Equation 1.6 to the participants at level b_2 (or use Equation 3.1 with $v_1 = 0$ and $v_2 = 1$).

A confidence interval for the simple main effect of B at a_1 is obtained by computing a linear contrast score using $q_1 = 1$ and $q_2 = 0$ and applying Equation 2.1 or Equation 3.1 with $v_1 = 1$ and $v_2 = -1$. Likewise, a confidence interval for the simple main effect of B at a_2 is obtained by computing a linear contrast score using $q_1 = 0$ and $q_2 = 1$ and applying Equation 2.1 or Equation 3.1 with $v_1 = 1$ and $v_2 = -1$.

The procedures described above suggest a general approach for computing confidence intervals for the effects in a 2 × 2 mixed design. The basic idea is to compute an appropriate linear contrast score for the within-subject factor and then estimate an appropriate function of population means for the between-subjects factor. The following table summarizes the coefficients that define the effects in a 2 × 2 mixed design where the v_j coefficients are applied to the levels of the between-subjects factor (Factor A) and the q_j coefficients define a linear contrast score for the within-subjects factor (Factor B).

v_1	v_2	q_1	q_2	Effect
1/2	1/2	1	-1	Main effect of A
1	-1	1/2	1/2	Main effect of B
1	-1	1	-1	AB interaction effect
1	-1	1	0	Simple main effect of B at a_1
1	-1	0	1	Simple main effect of B at a_2
1	0	1	-1	Simple main effect of A at b_1
0	1	1	-1	Simple main effect of A at b_2

Alternatively, the `ci.2x2.mean.mixed` and `ci.2x2.stdmean.mixed` functions in the `statpsych` package will compute confidence intervals for the AB interaction, the two main effects, and the four simple main effects in a 2 × 2 mixed design.

This approach can be used to estimate a wide range of interesting effects in a general $a \times b$ mixed design. Consider a 3 × 4 mixed design where the between-subject factor (Factor A) has 3 levels and the within-subjects factor (Factor B) has 4 levels. A pairwise main effect comparing levels 1 and 2 of Factor B would use $v_1 = 1/3$, $v_2 = 1/3$, $v_3 = 1/3$, $q_1 = 1$, $q_2 = -1$, $q_3 = 0$, and $q_4 = 0$. A pairwise main effect comparing levels 2 and 3 of Factor A would use $v_1 = 0$, $v_2 = 1$, $v_3 = -1$, $q_1 = 1/4$, $q_2 = 1/4$, $q_3 = 1/4$, and $q_4 = 1/4$. A pairwise interaction effect comparing levels 1 and 2 of Factor A and levels 1 and 2 of Factor B would use $v_1 = 1$, $v_2 = -1$, $v_3 = 0$, $q_1 = 1$, $q_2 = -1$, $q_3 = 0$, and $q_4 = 0$. A Factor B main effect contrast that compares the average of levels 1 and 2 with level 3 would use $v_1 = 1/3$, $v_2 = 1/3$, $v_3 = 1/3$, $q_1 = 1/2$, $q_2 = 1/2$, $q_3 = -1$, and $q_4 = 0$.

4.17 Two-way Analysis of Variance for Mixed Designs

The ANOVA table for a two-factor mixed design, where Factor B is a between-subjects factor and Factor A is a within-subjects factor, is shown below where n is the total sample size. The notation S(B) indicates that Factor S, the random "Subject" factor, is *nested* within the levels of Factor B (the between-subjects factor). One factor is said to be nested within another factor if only some levels of the first factor occur within each level of the second factor (e.g., one group of participants are in level 1 of Factor B and another group of participants are in level 2 of Factor B). The notation AS(B) indicates that the combination of within-subjects factor levels (Factor A) and the levels of Factor S are nested within the levels of the between-subjects factor (Factor B).

Source	*SS*	*df*	*MS*	*F*
A	SS_A	$df_A = a - 1$	$MS_A = SS_A/df_A$	$MS_A/MS_{AS(B)}$
B	SS_B	$df_B = b - 1$	$MS_B = SS_B/df_B$	$MS_B/MS_{A(B)}$
S(B)	$SS_{S(B)}$	$df_{S(B)} = n - b$	$MS_{S(B)} = SS_{S(B)}/df_{S(B)}$	
AB	SS_{AB}	$df_{AB} = (a - 1)(b - 1)$	$MS_{AB} = SS_{AB}/df_{AB}$	$MS_{AB}/MS_{AS(B)}$
AS(B)	$SS_{AS(B)}$	$df_{AS(B)} = (a - 1)(n - b)$	$MS_{AS(B)} = SS_{AS(B)}/df_{AS(B)}$	

The following generalized eta-squared estimates can be computed from the ANOVA table for a two-way mixed design. Equation 3.19 can be used to reduce the positive bias of these estimates.

$$\hat{\eta}_A^2 = SS_A/(SS_A + SS_{S(B)} + SS_{AS(B)}) \quad (4.17a)$$

$$\hat{\eta}_B^2 = SS_B/(SS_B + SS_{S(B)} + SS_{AS(B)}) \quad (4.17b)$$

$$\hat{\eta}_{AB}^2 = SS_{AB}/(SS_{AB} + SS_{S(B)} + SS_{AS(B)}) \quad (4.17c)$$

A two-way ANOVA for a mixed design can be used to test three omnibus null hypotheses: a zero main effect of Factor A, a zero main effect of Factor B, and a zero AB interaction effect. Hypothesis testing in the two-way ANOVA for a mixed design suffers from the same problems as the two-way between-subjects ANOVA. Confidence intervals for standardized or unstandardized linear contrasts are recommended alternatives or supplements to the *F* tests. The same analysis strategy for a two-factor between-subjects design that was described in section 3.13 also applies to the two-factor mixed design.

4.18 Counterbalancing

The usefulness of a within-subject experimental design is limited by the assumption of no practice, fatigue, or carryover effects. It is possible to completely control for practice, fatigue, and a specific type of carryover effect by *counterbalancing* the order of the treatment conditions. For example, with $a = 2$ treatment conditions, one group of participants receives treatment a_1 followed by treatment a_2 (a_1 -> a_2) and a second group receives treatment a_2 followed by treatment a_1 (a_2 -> a_1). A design that uses all possible order conditions is called a *completely counterbalanced design*. An equal number of participants should be assigned to each order condition. With $a = 3$ treatments, the six possible orders are given below.

a_1 -> a_2 -> a_3	a_2 -> a_3 -> a_1
a_1 -> a_3 -> a_2	a_3 -> a_1 -> a_2
a_2 -> a_1 -> a_3	a_3 -> a_2 -> a_1

With $a = 4$ treatments there are 24 possible order conditions. This requires a sample size that is a multiple of 24 which could be difficult to achieve. Instead of using all possible orders, a particular subset of the 24 possible orders can used to control for practice, fatigue, and a specific type of carryover effect. With $a = 4$ treatments, the following four order conditions are recommended.

a_1 -> a_2 -> a_4 -> a_3
a_2 -> a_3 -> a_1 -> a_4
a_3 -> a_4 -> a_2 -> a_1
a_4 -> a_1 -> a_3 -> a_2.

This particular subset of order conditions is called *balanced Latin Square* (BLS) counterbalancing in which each treatment condition immediately follows every other treatment condition in only one (for a even) or two (for a odd) of the order conditions. Using the above four order conditions, the sample size can be a multiple of 4 rather than 24. BLS counterbalancing requires only a order conditions for even values of a but $2a$ order conditions are needed for odd values of a. With $a = 5$ treatments, the following ten order conditions provide BLS counterbalancing.

a_1 -> a_2 -> a_5 -> a_3 -> a_4
a_2 -> a_3 -> a_1 -> a_4 -> a_5
a_3 -> a_4 -> a_2 -> a_5 -> a_1
a_4 -> a_5 -> a_3 -> a_1 -> a_2
a_5 -> a_1 -> a_4 -> a_2 -> a_3

a_4 -> a_3 -> a_5 -> a_2 -> a_1
a_5 -> a_4 -> a_1 -> a_3 -> a_2
a_1 -> a_5 -> a_2 -> a_4 -> a_3
a_2 -> a_1 -> a_3 -> a_5-> a_4
a_3 -> a_2 -> a_4 -> a_1-> a_5.

To illustrate how counterbalancing can control for practice, fatigue, and a specific type of carryover effect, consider a 2 × 2 mixed design where the within-subjects factor (Factor A) has two treatment conditions (a_1 and a_2) and the between-subjects factor (Factor B) has two order conditions ($b_1 = a_1$ -> a_2 and $b_2 = a_2$ -> a_1). A carryover effect can be *symmetric* or *asymmetric*. With a symmetric carryover effect, the carryover from treatment a_1 to a_2 is equal to the carryover from treatment a_2 to a_1; otherwise, the carryover effect is asymmetric.

Let ζ be the unknown value of a practice effect, a fatigue effect, or a symmetric carryover effect. The population means for the 2 × 2 mixed design are given below.

		Factor B (Order) b_1	b_2
Factor A (Treatment)	a_1	μ_1	$\mu_1 + \zeta$
	a_2	$\mu_2 + \zeta$	μ_2

The main effect of Factor A is equal to $(\mu_1 + \mu_1 + \zeta)/2 - (\mu_2 + \zeta + \mu_2)/2 = \mu_1 - \mu_2$ and the effects of practice, fatigue, or symmetric carryover has been eliminated by counterbalancing. Recall from Chapter 2 that in a two-group experiment the population mean of the response variable under the a_1 treatment is equal to μ_1, and the population mean of the response variable under the a_2 treatment is equal to μ_2. Thus, the main effect of Factor A in this 2 × 2 mixed design is identical to the effect that would be estimated using a two-group experimental design. The AB interaction is $[\mu_1 - (\mu_1 + \zeta)] - [(\mu_2 + \zeta) - \mu_2] = -2\zeta$ which describes the magnitude of the carryover effect.

In this 2 × 2 mixed design, we do not follow the convention of analyzing simple main effects if an AB interaction effect is detected because the simple main effects of Factor A in this design are biased estimates of $\mu_1 - \mu_2$. The main effect of Factor A should always be examined in this type of design regardless of the size of the AB interaction.

With asymmetric carryover, the main effect of Factor A will not equal $\mu_1 - \mu_2$. Let ζ_1 be the carryover from a_1 to a_2 and let ζ_2 be the carryover from a_2 to a_1. With asymmetric carryover, the main effect of A is equal to $(\mu_1 + \mu_1 + \zeta_1)/2 - (\mu_2 + \mu_2 + \zeta_2)/2 = \mu_1 - \mu_2 + (\zeta_1 - \zeta_2)/2$ and the estimate of $\mu_1 - \mu_2$ is biased by an amount equal to $(\zeta_1 - \zeta_2)/2$. Thus, complete counterbalancing will not control for asymmetric carryover. The main effect of Factor B (the order factor), which is equal to $(\mu_1 + \mu_2 + \zeta_2)/2 - (\mu_1 + \mu_2 + \zeta_1)/2 = (\zeta_2 - \zeta_1)/2$, describes the degree of carryover asymmetry.

Complete and BLS counterbalanced designs control for symmetric carryover and any pattern of practice or fatigue effects. For example, the practice effect could be large after the first treatment but smaller for subsequent treatments, or there might not be any fatigue effect until the last treatment. Complete and BLS counterbalancing also controls for some specific patterns of asymmetric carryover

effects. For example, with $a = 3$ the following table describes all possible carryover effects from one treatment to the treatment that immediately follows.

		Carryover from: a_1	a_2	a_3
To:	a_1	---	ζ_3	ζ_5
	a_2	ζ_1	---	ζ_6
	a_3	ζ_2	ζ_4	---

If it is reasonable to assume that $\zeta_3 + \zeta_5 = \zeta_1 + \zeta_6 = \zeta_2 + \zeta_4$, then estimates of all pairwise differences among the $a = 3$ treatment means will be unbiased in a mixed design with complete or BLS counterbalancing.

In studies that use a within-subjects treatment factor, complete or BLS counterbalancing of the treatment orders is recommended unless there is a compelling argument that there will be no practice, fatigue, or carryover effects. When complete or BLS counterbalancing is used, the data should be analyzed using a mixed design rather than a single factor within-subjects design. The mixed design will provide more powerful tests and narrower confidence intervals of the treatment effects than the statistical methods for a single factor within-subject design. The lack of power and precision in the one-factor design is due to an addition of a practice, fatigue, or carryover effect to only some of the participant scores within each treatment condition. This increases the variance of the response variable within each treatment condition and also decreases the correlation between any two treatment conditions. The increased variance and decreased correlation produce a larger standard error, which in turn results in less powerful tests and wider confidence intervals.

Complete and BLS counterbalancing only controls for specific types of asymmetric carryover effects. It might be possible to reduce all types of carryover effects by increasing the length of time between treatments or requiring participants to complete some unrelated task between treatments. If an unknown pattern of asymmetric carryover effects is possible, it would be prudent to use a between-subjects treatment factor rather than a within-subjects treatment factor.

4.19 Reliability Designs

In the physical sciences, attributes such as weight, length, time, volume, and pressure can be measured with great accuracy. When measuring the weight of an object, two laboratory-grade scales will yield virtually the same value, two different technicians using the same scale will obtain virtually the same value, or the same technician using the same scale will obtain virtually the same value on two different occasions. In the behavioral sciences, however, psychological attributes cannot be measured with high accuracy. For example, if a particular student takes two forms of the ACT, or takes the same form of the ACT on two different occasions, or if two expert graders both evaluate the student's written essay, the two scores could be substantially different.

Measurement error for person i is the unknown and unpredictable difference between that person's true score for some attribute and a measurement of the attribute for that person. For person i, the measured attribute score is y_i, the true score is T_i and the measurement error is ϵ_i. In any given study population, the variance of the observed measurements σ_y^2 is assumed to equal the variance of the true scores (σ_T^2) plus the variance of the measurement errors (σ_ϵ^2). The *reliability coefficient* is defined as the true score variance divided by the observed score variance σ_T^2/σ_y^2 and has a range of 0 to 1. A reliability coefficient of 1 indicates that the measurements contain no measurement error and a reliability coefficient of 0 indicates that the measurements are pure measurement error.

The above definition of reliability requires knowledge of the true score variance, but when measuring a psychological attribute, person's true score will be unknown. A fundamental theorem in psychometrics shows that the reliability of a measurement can be estimated using multiple measurements of the same attribute. In a reliability design where $a \geq 2$ equally reliable measurements are obtained from a random sample of n participants, a one-way within-subjects ANOVA can be used to estimate the reliability of any single measurement where the levels of Factor A represent the multiple measurements. The multiple measurements per participant could be ratings from $a \geq 2$ different raters (to estimate *interrater reliability*), scores on a particular questionnaire at $a = 2$ points in time (to estimate *test-retest reliability*), scores from $a \geq 2$ different forms of a test or

questionnaire (to estimate *alternate form reliability*), or the responses to $a \geq 2$ quantitatively scored items of a questionnaire (to estimate *internal consistency reliability*).

If two or more measurements of an attribute are obtained from a sample of n participants, the following estimate of the reliability of a single measurement (denoted as ρ_1) can be obtained from a one-way within-subjects ANOVA table.

$$\hat{\rho}_1 = \frac{MS_S - MS_{AS}}{MS_S + (a-1)MS_{AS}} \tag{4.18}$$

The reliability of a sum (or average) of $a \geq 2$ equally reliable measures (denoted as ρ_a) is estimated as

$$\hat{\rho}_a = 1 - \frac{MS_{AS}}{MS_S} \tag{4.19}$$

and is referred to as *coefficient alpha* (or *Cronbach's alpha*).

The reliability of the sum (or average) of $a \geq 2$ equally reliable measures will be more reliable than any single measurement. The following two *Spearman-Brown* formulas show the relation between ρ_a and ρ_1

$$\rho_a = \frac{a\rho_1}{1 + (a-1)\rho_1} \tag{4.20}$$

$$\rho_1 = \frac{\rho_a}{a - (a-1)\rho_a} \tag{4.21}$$

The value of a in Equations 4.20 and 4.21 can represent the number of individual measurements or the number of sets of measurements. For example, $a = 1$ could represent one set of 20 reaction time trials that are averaged to define the response variable. Setting $a = 2$ in Equation 4.20 would give the reliability of the average reaction time score based on 40 trails. The `spearmanbrown` function in the `statpsych` package will compute Equation 4.20.

To illustrate the use of the Spearman-Brown formulas, suppose the reliability of a single measurement is 0.5, then the reliability of the sum or average of $a = 3$ equally reliable measurements is 3(0.5)/[1 + 2(0.5)] = 0.75. Or suppose the reliability of a 5-item questionnaire score has a reliability of 0.9. Assuming equally reliable items, the reliability of a single item is 0.9/[5 – 4(0.9)] = 0.643.

When the measurements represent different test forms, the reliability of a single measurement (ρ_1) is usually of primary interest. When the measurements represent the items of a questionnaire, the reliability of the sum or average (ρ_a) of the multiple item scores will be of primary interest.

An approximate $100(1-\alpha)$% confidence interval for ρ_a is

$$1 - exp[ln(1-\hat{\rho}_a) - ln\{n/(n-1)\} \pm z_{\alpha/2}\sqrt{\frac{2a}{(a-1)(n-2)}}\,] \qquad (4.22)$$

and a $100(1-\alpha)$% confidence interval for ρ_1 is obtained by transforming the endpoints of the confidence interval for ρ_a using Equation 4.21. An exact confidence interval for ρ_a can be computed in SPSS and the `ci.cronbach` function in the `statpsych` package. Equation 4.22 is useful for other types of reliability coefficients that are described in Chapter 2 of Volume 4 which do not have exact confidence interval formulas.

Example 4.8. Two parole officers (a = 2) independently assigned recidivism scores to a random sample of 50 sex offenders taken from a Midwest prison population of about 16,000 sex offenders. The estimate of ρ_2 was 0.87. An approximate 95% confidence interval for ρ_2 is [0.77, 0.93]. The researcher is 95% confident that the reliability of the average of these two parole officer ratings in the population of sex offenders is between 0.77 and 0.93. A 95% confidence interval for the reliability of a single parole officer rating is [0.62, 0.87].

4.20 Effects of Measurement Error

In between-subjects and within-subjects designs, measurement error increases the variance of the response variable within treatment conditions, which reduces the power of a statistical test and increases the width of a confidence interval. In within-subject designs, measurement error has the additional detrimental effect of attenuating the correlations among the measurements, which contributes to a further decrease in power and an increase in confidence interval width. An important consequence of measurement error in a between-subject or within-subject design is the need for a larger sample size. Measurement error also attenuates the estimates of standardized effect size measures (e.g., δ, φ, and η^2). Effect size attenuation is discussed in section 1.30 of Volume 2.

The sample size requirement can be reduced if the reliability of the response variable can be improved. Some psychological scales have both short-form and long-form versions. The long-form version will be more reliable than the short-form version and a smaller sample size can be used with the long-form version. Using the short-form rather than the more reliable long-form has the benefit of requiring less testing time but will requires a larger sample size.

The reliability of the response variable can be increased by using an average of two or more equally reliable measurements of the attribute under investigation (see Equation 4.20). Multiple measurements of an attribute (referred to here as "measurement trials") are commonly used in studies that measure response speed, physical performance, and various types of brain activities. Increasing the number of measurement trials has the benefit of increasing the reliability of the response variable, which in turn increases hypothesis testing power and reduces confidence interval width. Alternatively, increasing the number of measurement trials increases testing time and can have a negative effect on the motivation and attention of the participants.

When designing a study to achieve desired power or confidence interval precision, the researcher can choose to use a less reliable response variable and take a larger sample size or use a more reliable response variable and take a smaller sample size. If the cost of a treatment is high or if certain types of participants are difficult to obtain, the total cost of the study could be reduced by using a more costly but more reliable response variable. In a within-subjects design that uses the average of multiple measurements as the response variable, using fewer measurement trials will require a larger sample size, but it may then be possible to test each participant under more conditions. For example, instead of using 40 measurement trials in a within-subjects experiment with two treatment conditions, four treatment conditions could be assessed using 20 measurement trials while maintaining the total testing time for each participant. However, with 20 measurement trials, a larger sample size will be needed to maintain the same power and confidence interval width that would be obtained using 40 measurement trials.

4.21 Assumptions

In addition to the assumptions of random sampling and independence among participants, the paired-samples *t*-test assumes that the difference scores have an approximate normal distribution in the study population. For within-subjects designs with three or more levels, the within-subjects ANOVA *F* test requires two assumptions in addition to random sampling and independence among participants: the population variances of all pairs of difference scores are assumed to be equal (the *sphericity* assumption), and all pairs of difference scores are assumed to have an approximate normal distribution in the study population.

The within-subjects ANOVA *F* test will not perform properly when the sphericity assumption has been violated, even in large samples. Furthermore, the traditional diagnostic tests for sphericity are not helpful because they frequently fail to detect a serious assumption violation. If the number of participants is greater than the number of within-subject conditions, a *multivariate test* is the recommended alternative to the within-subjects ANOVA *F* test. The multivariate test does not require sphericity. The within-subjects ANOVA *F* test and the multivariate test are sensitive to skewness of the difference scores, but these tests will perform properly unless the difference scores are highly skewed and the sample size is small ($n < 20$). Greater amounts of skewness can be tolerated with larger sample sizes.

The confidence interval for a linear contrast of population means (Equation 4.7 or Equation 4.1 for the special case of a mean difference) requires only one assumption in addition to the random sampling and independence assumptions. The only additional assumption is that the linear contrast scores have an approximate normal distribution in the study population. Skewness, rather than kurtosis, of the linear contrast scores is the major concern. A $100(1 - \alpha)\%$ confidence interval for a linear contrast of population means will have a coverage probability close to $1 - \alpha$ unless the linear contrast scores are more than moderately skewed and the sample size is small ($n < 20$). Greater amounts of skewness can be tolerated with larger sample sizes.

If the scores at each level of the within-subject factor have similar variability and skewness, then the linear contrast scores should have very little skewness. The `sim.ci.mean.ps` function in the `statpsych` package can be used to assess the

performance of the confidence interval for $\mu_1 - \mu_2$ in a paired-samples design with different sample sizes, different values of σ_1/σ_2, different correlations, and different distribution shapes of the y_1 and y_2 scores.

The confidence intervals for δ, δ'', and φ are very sensitive to a violation of the normality assumption regardless of sample size. However, the sensitivity to moderate nonnormality is minimal for small absolute values of δ, δ'', or φ (less than .5) when the population variances are similar. Consequently, the normality assumption is not a concern when a confidence interval for δ or δ'' is used in an equivalence test where the standardized mean difference is assumed to be small. The `sim.ci.stdmean.ps` function in the `statpsych` package can be used to assess the performance of the confidence intervals for δ and δ'' in a paired-samples design with different sample sizes, different values of σ_1/σ_2, different correlations, and different distribution shapes of the y_1 and y_2 scores.

A $100(1 - \alpha)$% confidence interval for a ratio of means, which is a unitless measure of effect size for ratio scale data, will have a true coverage probability close to $1 - \alpha$ unless the sample size is small and the y scores are more than moderately nonnormal.

In addition to the random sampling, independence, and normality assumptions, the two-way mixed ANOVA F tests for the within-subject factor and the interaction assume sphericity within each level of the between-subjects factor and also assume that the within-subjects variances and correlations are equal across the levels of the between-subjects factor. Hypothesis tests in the mixed ANOVA will not perform properly when the sphericity assumption or the assumption of equal variances and correlations across groups has been violated, even in large samples. A multivariate test is recommended for mixed designs. The multivariate test does not assume sphericity but it does assume that the within-subjects variances and covariances are equal across groups.

The normality assumption in a mixed design applies to the response scores that are used to assess a particular effect. For example, in a 2 × 2 mixed sign, tests and confidence intervals for the AB interaction and the within-subjects main effect assume that the within-subjects difference scores have an approximate normal distribution, while tests and confidence intervals for the between-subjects main

effect assume that the average within-subjects scores have an approximate normal distribution.

4.22 Missing Data

Missing data is more of a problem in a within-subjects ANOVA than a between-subjects ANOVA. If a participant fails to produce a score for *any* of the within-subject conditions, statistical packages will eliminate that participant from the analysis. This is called *listwise deletion*. Missing data is less of a problem with pairwise comparisons because this analysis only needs to drop participants who do not have the two scores required for a particular pairwise comparison. This is called *pairwise deletion*. As in between-subjects designs, a random loss of data does not affect the internal or external validity of a within-subjects study, but it will decrease the power of statistical tests and increase confidence interval widths. If a nonrandom subset of participants has one or more missing scores and must be dropped from the analysis, then the remaining participants are no longer a random sample from the specified study population.

4.23 Distribution-free Methods

If the response variable is skewed, a difference in medians $(\theta_1 - \theta_2)$ could be a more meaningful measure of effect size than a difference in means. An approximate $100(1-\alpha)\%$ confidence interval for $\theta_1 - \theta_2$ in a paired-samples design is

$$\hat{\theta}_1 - \hat{\theta}_2 \pm z_{\alpha/2}\sqrt{SE_{\hat{\theta}_1}^2 + SE_{\hat{\theta}_2}^2 - 2(4p_{00} - 1)SE_{\hat{\theta}_1}SE_{\hat{\theta}_2}} \tag{4.23}$$

where $SE_{\hat{\theta}_j}^2$ is defined in Equation 1.10 and p_{00} is the sample proportion of paired observations for which $y_1 < \hat{\theta}_1$ and $y_2 < \hat{\theta}_2$. Equation 4.22 assumes random sampling, independence among participants, and that tied scores are rare. Equation 4.23 can be used to perform a directional two-sided test, an equivalence test, a noninferiority test, or a superiority test. The `ci.median.ps` function in the `statpsych` package will compute Equation 4.23. The `ci.2x2.median.ws` in the `statpsych` package will compute confidence intervals for the AB interaction, the two main effects, and the four simple main effects in terms of medians in a 2 × 2 within-subjects design.

The `sim.ci.median.ps` function in the `statpsych` package can be used to assess the performance of the confidence interval for $\theta_1 - \theta_2$ in a paired-samples design with different sample sizes, different values of σ_1/σ_2, different correlations, and different distribution shapes of the y_1 and y_2 scores. It is instructive to compare these results with those from `sim.mean.ps`.

If the scale of the response variable is not familiar to the intended audience, a unitless measure of effect size is the proportion of people in the study population who have y_1 scores that are greater than their y_2 scores. The population proportion is denoted as π and is estimated as $\hat{\pi} = (f + 2)/(n + 4)$ where f is the number of participants in the sample who have y_1 scores that are greater than their y_2 scores. Equation 1.11 can be used to obtain an approximate confidence interval for π. The `ci.prop` function in the `statpsych` package can be used to compute Equation 1.11. A confidence interval of π is useful in applications where the difference scores ($d_i = y_{1i} - y_{2i}$) are highly nonnormal.

Example 4.9. Ten participants were randomly selected from a study population of about 1,200 college-bound high school seniors and were paid to record their hours of study during the month of April and paid again one year later (also during the month of April) at college. Their average weekly hours of study are shown below.

Student:	1	2	3	4	5	6	7	8	9	10
High School	4.1	6.9	12.2	9.7	7.6	11.5	29.8	10.5	3.0	24.2
College	8.3	14.0	16.7	10.5	15.0	20.4	37.9	18.2	7.8	32.5

The 95% confidence interval for $\theta_1 - \theta_2$ is [-10.12, -1.38]. In this example, the 95% confidence interval for $\mu_1 - \mu_2$ is [-14.93, 2.57], which includes 0 and is wider than the confidence interval for $\theta_1 - \theta_2$. (*continued*)
A confidence interval for $\theta_1 - \theta_2$ can be narrower than a confidence interval for $\mu_1 - \mu_2$ when the scores are highly skewed or a few scores have extreme values.

If the response variable is measured on a ratio scale and the response variable is skewed, a ratio of population medians (θ_1/θ_2) could be a more meaningful measure of effect size than a ratio of means. Like a ratio of means, a ratio of medians is a unitless measure of effect size. The `ci.ratio.median.ps` function in the `statpsych` package will compute a confidence interval for θ_1/θ_2.

Example 4.10. In *Example 4.2*, a face drawing error score was measured under an upright and inverted condition in a within-subjects design. The 95% confidence interval for a ratio of population median drawing accuracy scores is [1.06, 1.36]. This result indicates that the population median drawing error is 1.06 to 1.36 times as large for an inverted face as an upright face.

For a paired-samples design, the sign test of Chapter 1 can be applied to the difference scores to test H_0: $\theta = 0$ where θ is the population median of the difference scores. The *Wilcoxon signed rank* test is a more powerful test of H_0: $\theta = 0$ than the sign test and assumes that the distribution of difference scores is symmetric. The Wilcoxon signed rank test is usually a little less powerful than the paired-samples *t*-test, but it can be more powerful than the *t*-test if the response variable is highly leptokurtic.

For within-subjects designs with $a > 2$, the *Friedman test* is a distribution-free alternative to the *F* test in the one-way within-subjects ANOVA. Although the Friedman test might be preferred to the *F* test in situations where the response variable is skewed and the sample size is small, it does not provide useful scientific information because the null hypothesis is known to be false in virtually every study. An alternative to the Friedman test involves performing Wilcoxon signed rank tests for some or all pairwise comparisons using the Holm procedure.

4.24 Variability Assessment

Variability assessment is just as important in within-subjects designs as it is in between-subjects designs. In addition to hypothesis tests or confidence intervals for a difference in population means, population medians, or standardized means, it is informative to also examine confidence intervals for τ_j (using Equation 1.14) and τ_1/τ_2 where τ_j is the population MAD under condition *j*. In a within-subjects experiment, a treatment that improves the mean of the response variable and also reduces the variability of the response variable is usually more beneficial than a treatment that only improves the mean of the response variable.

An approximate 100(1 – α)% confidence interval for τ_1/τ_2 in a paired-samples design is

$$exp[ln(\hat{\tau}_1/\hat{\tau}_2) \pm z_{\alpha/2}SE_{ln(\hat{\tau}_1/\hat{\tau}_2)}] \tag{4.24}$$

where $SE_{ln(\hat{\tau}_1/\hat{\tau}_2)} = \sqrt{(V_1 + V_2 - 2\hat{\rho}\sqrt{V_1 V_2})/n}$, $V_1 = (\hat{\mu}_1 - \hat{\theta}_1)^2/\hat{\tau}_1^2 + \hat{\sigma}_1^2/\hat{\tau}_1^2 - 1$, $V_2 = \left(\hat{\mu}_2 - \hat{\theta}_2\right)^2/\hat{\tau}_2^2 + \hat{\sigma}_2^2/\hat{\tau}_2^2 - 1$, and $\hat{\rho}_{d_1 d_2}$ is the estimated Pearson correlation between the $|y_{1i} - \hat{\theta}_1|$ and $|y_{2i} - \hat{\theta}_2|$ absolute deviation scores. Equation 4.24 assumes the y scores have an approximate normal distribution in the study population, but this assumption is not a concern if $n \geq 30$ and the population y scores are not extremely nonnormal within each condition. The `ci.mad2.ps` function in the `statpsych` package will compute Equation 4.24.

Example 4.11. A high school in Los Angeles has about 900 students who are recent immigrants from Central America. Eighty of these students were randomly sampled and given six months of a new type of English language instruction. The participants were given an English language proficiency test (scored on a 0 to 70 scale) before and after training. The 95% confidence interval for $\mu_1 - \mu_2$ is [-9.33, -6.44], and the 95% confidence interval for τ_1/τ_2 is [1.47, 2.70]. These results indicate that the new type of instruction will improve the population mean English language proficiency. The confidence interval for τ_1/τ_2 indicates that the population MAD prior to training would be 1.47 to 2.70 times as large as the population MAD after training. Thus, the new training would not only increase the population mean of the proficiency scores but also would reduce the population variability in proficiency scores.

The variability of pretest and difference scores in a pretest-posttest design provides useful information. Equation 1.14 can be used to obtain a confidence interval for the population MAD pretest and difference scores. Large variability of pretest or difference scores indicates that additional research might be needed to discover a classification factor that explains some of the variability in the pretest or difference scores. If such a classification factor can be found, this factor could then be used with the pretest-posttest factor in a two-factor mixed design. The two-factor mixed design could provide important information about the main effect of the classification factor and the interaction effect.

4.25 Using Prior Information

Suppose the results of a paired-samples study has been reported in the literature, but the confidence interval for the population mean difference was wide. This study could be replicated in a new study population in an effort to assess replicability and obtain a narrower confidence interval for the population mean

difference that describes both the prior study population and the follow-up study population. If the sample means, sample standard deviations, sample correlation, and sample size have been reported in the prior study, the `replicate.mean.ps` function in the `vcmeta` package will compute confidence intervals for μ_{d1}, μ_{d2}, $\mu_{d1} - \mu_{d2}$, and $(\mu_{d1} + \mu_{d2})/2$ where μ_{dj} is the population mean difference in study j. This function requires an estimate of the Pearson correlation between the paired observations in each study. Most studies that use a paired-samples design do not report the Pearson correlation but it can be recovered if the sample means, sample standard deviations, sample size, and paired-sample t-test are reported. The `cor.from.t` function in the `vcmeta` package will perform this computation.

Example 4.12. A study in San Luis Obispo used a paired-samples design with a random sample of 20 registered voters to compare the favorability ratings of two proposals to assist homeless teenagers. This study was replicated in Santa Cruz using a random sample of 80 registered voters. The summary results for these two studies are given below.

Study	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\rho}_{12}$	n
1	14.37	13.82	1.14	1.31	.672	20
2	14.52	12.47	1.39	1.68	.629	80

The 95% confidence interval for μ_{d1} was [0.08, 1.02]. The 95% confidence interval for $(\mu_{d1} + \mu_{d2})/2$ is [1.02, 1.57]. We can be 95% confident that the mean favorability rating, averaged over the two study populations, is 1.02 to 1.57 points greater for proposal 1 than proposal 2.

As explained in section 1.32, the Bayesian approach uses a subjective prior distribution rather than the actual results from a prior study. Equations 1.21 and 1.22 can be applied to μ_l (with μ_d as a special case) in a within-subjects design by replacing $\tilde{\mu}$ and $\sigma_{\tilde{\mu}}$ with a subjective mean and standard deviation of a prior distribution of plausible values of μ_l. The use of subjective prior distributions is rare in psychological research where objectivity is the primary focus. However, the use of a skeptical prior distribution (see section 3.23) should be welcomed in scientific research. A skeptical prior distribution assumes that the effect (e.g., mean difference, standardized mean difference, linear contrast of means, standardized linear contrast of means, median difference, linear contrast of medians) is likely to be small in most psychological studies. For example, a skeptic might be 95% confident that a population standardized main effect in a 2 x 2 within-subjects

design will be between -.2 and .2. This skeptical prior information can be characterized by a normal distribution with a mean of 0 and a standard deviation of .2/1.96 = .102. Suppose the sample standardized main effect is .92 with a standard error of .16 (95% CI [.62, 1.24]). Applying the `ci.bayes.normal` function gives a Bayesian estimate of the standardized main effect of .27 with a 95% credible interval of [.10, .43]. The Bayesian results are more modest but are also more likely to replicate in a future study.

4.26 Sample Size Requirement for Desired Precision

The width of the confidence interval for $\mu_1 - \mu_2$ in a paired-samples design depends on the correlation between the two measurements with a larger correlation producing a narrower width. The required sample size to estimate $\mu_1 - \mu_2$ with desired precision and desired confidence interval width (w) in a paired-samples design is approximately

$$n = 8\tilde{\sigma}^2(1 - \tilde{\rho}_{12})(\frac{z_{\alpha/2}}{w})^2 + \frac{z_{\alpha/2}^2}{2} \quad (4.25)$$

where $\tilde{\rho}_{12}$ is a planning value of the Pearson correlation between the two measurements, and $\tilde{\sigma}^2$ is a planning value of the average within-condition variance. Note that the sample size requirement is larger for smaller values of $\tilde{\rho}_{12}$. Using the smallest likely correlation planning value will give a conservatively large sample size requirement. The `size.ci.mean.ps` function in the `statpsych` package computes Equation 4.25.

Example 4.13. A researcher wants to compare the opinions of husbands and wives regarding the inclusion of sexual stereotype issues in elementary school curricula. The researcher wants to estimate $\mu_1 - \mu_2$ with 95% confidence and wants the width of the interval to be about 2. From previous research, the researcher decides to set $\tilde{\sigma}^2$ = 5.0 and $\tilde{\rho}_{12}$ = .5. To estimate $\mu_1 - \mu_2$ using a paired-samples study of husbands and wives, the required number of couples to sample is approximately $8(5.0)(1 - .5)(1.96/2)^2 + 1.92 = 21.1 \approx 22$.

The approximate sample size required to estimate δ in a paired-samples design with desired confidence and desired confidence interval width is

$$n = [\tilde{\delta}^2(1 + \tilde{\rho}_{12}^2) + 8(1 - \tilde{\rho}_{12})](\frac{z_{\alpha/2}}{w})^2 \quad (4.26)$$

where $\tilde{\delta}$ is a planning value of δ. The approximate sample size required to estimate δ" in a paired-samples design with desired confidence and desired confidence interval width is

$$n = [2\tilde{\delta}^{"2} + 8(1 - \tilde{\rho}_{12})](\frac{z_{\alpha/2}}{w})^2 \quad (4.27)$$

where $\tilde{\delta}$" is a planning value of δ". The `size.ci.stdmean.ps` function in the `statpsych` package computes Equations 4.26 and 4.27. Using the largest likely value for the standardized mean difference and the smallest likely value for $\tilde{\rho}_{12}^2$ will give a conservatively large sample size requirement.

If the researcher plans to report a confidence interval for $\mu_1 - \mu_2$ and finds it easier to specify the desired confidence interval width for δ, then multiplying the desired confidence interval width for δ by $\sqrt{\tilde{\sigma}^2}$ gives the corresponding desired confidence interval width for $\mu_1 - \mu_2$ that can be used in Equation 4.25.

The approximate sample size required to estimate μ_1/μ_2 in a paired-samples design with desired confidence and precision is

$$n = 8\tilde{\sigma}^2\left(\frac{1}{\tilde{\mu}_1^2} + \frac{1}{\tilde{\mu}_2^2} - \frac{2\tilde{\rho}_{12}}{\tilde{\mu}_1\tilde{\mu}_2}\right)[\frac{z_{\alpha/2}}{ln(\tilde{U}/\tilde{L})}]^2 + \frac{z_{\alpha/2}^2}{2} \quad (4.28)$$

where $\tilde{\mu}_j$ is a planning value of μ_j, $\tilde{U}$ is the desired upper limit, and $\tilde{L}$ is the desired lower limit. Equation 4.28 requires an accurate planning value for each population mean which limits its usefulness. The `size.ci.ratio.mean.ps` function in the `statpsych` package computes Equation 4.28.

In a one-factor or factorial within-subjects design with a total of m within-subject levels, the approximate sample size requirement to estimate $\sum_{j=1}^{m} q_j\mu_j$ with desired confidence and desired confidence interval width in a within-subjects study is

$$n = 4\tilde{\sigma}^2(\sum_{j=1}^{m} q_j^2)(1 - \tilde{\rho})(z_{\alpha/2}/w)^2 + \frac{z_{\alpha/2}^2}{2} \quad (4.29)$$

where $\tilde{\sigma}^2$ is a planning value of the average within-condition variance, and $\tilde{\rho}$ is a planning value of the average correlation among all pairs of measurements. A Bonferroni adjustment to α in the critical z-value can be used when two or more

simultaneous confidence intervals are required. The `size.ci.lc.mean.ws` function in the `statpsych` package computes Equation 4.28.

Example 4.14. A researcher wants to replicate a published study that compared four graphical user interfaces in a within-subjects design. The researcher wants a 95% confidence interval for $(\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4)/2$ that has a width of about 4.0. Using the sample variances and correlations from the original study as planning values, the average variance was 161.9 and the average correlation was 0.77. The required number of participants is approximately $n = 4(161.9)(¼ + ¼ + ¼ + ¼)(1 - 0.77)(1.96/4.0)^2 + 1.92 = 37.7 \approx 38$.

The sample size required to estimate a standardized linear contrast of means (φ) with desired confidence and desired confidence interval width in a within-subjects design with m within-subject levels is approximately

$$n = [2\tilde{\varphi}^2[1 + (m - 1)\tilde{\rho}^2]/m + 4(1 - \tilde{\rho})\sum_{j=1}^{m} q_j^2](\frac{z_{\alpha/2}}{w})^2 \qquad (4.30)$$

where $\tilde{\varphi}$ is a planning value for φ and $\tilde{\rho}$ is a planning value for the average correlation among all pairs of measurements. A Bonferroni adjustment to α in the critical z-value can be used when two or more simultaneous confidence intervals will be computed. The `size.ci.lc.stdmean.ws` function in the `statpsych` package computes Equation 4.30. Equations 4.29 and 4.30 assume $\sum_{j=1}^{m} q_j = 0$ (unlike the sample size formulas for a linear contrast of means in a between-subjects design which do not require the sum of the coefficients to equal zero).

It is often easier to specify the desired width of a confidence interval for φ than a confidence interval for $\sum_{j=1}^{m} q_j\mu_j$. If the researcher plans to report a confidence interval for $\sum_{j=1}^{m} q_j\mu_j$ and finds it easier to specify the desired confidence interval width for φ, then multiplying the desired confidence interval width for φ by $\sqrt{\tilde{\sigma}^2}$ gives the corresponding desired confidence interval width for $\sum_{j=1}^{m} q_j\mu_j$ that can be used in Equation 4.29.

Example 4.15. A researcher wants to estimate φ in a 4-level within-subject experiment with 95% confidence and contrast coefficients 1/3, 1/3, 1/3, and -1. After reviewing previous research, the researcher decides to set $\tilde{\varphi} = 0.5$, $\tilde{\rho} = 0.7$, and $w = 0.4$. The required sample size is approximately $n = 4[0.25\{1 + 3(0.49)\}/8 + 0.3(1.33)](1.96/0.4)^2 = 45.8 \approx 46$.

The sample size required to estimate Cronbach's reliability (ρ_a) with desired confidence and desired confidence interval width (w) for an a-item questionnaire is approximately

$$n = 8[a/(a-1)](1-\tilde{\rho}_a)^2\,(\tfrac{z_{\alpha/2}}{w})^2 + 2 \qquad (4.31)$$

where $\tilde{\rho}_a$ is a planning value for ρ_a. The `size.ci.cronbach` function in the `statpsych` package uses Equation 4.31 as a first-step sample size approximation and then makes an adjustment to improve its accuracy. Using the smallest likely value for $\tilde{\rho}_a$ will give a conservatively large sample size requirement.

Example 4.16. A researcher wants to estimate Cronbach's reliability for a 5-item questionnaire with 95% confidence and a desired width of .15. The researcher set $\tilde{\rho}_a = .75$. The required sample size is approximately $n = 8[5/4](1 - .75)^2(1.96/0.15)^2 = 108.7 \approx 109$.

Specifying a narrow confidence interval width in Equations 4.25 - 4.31 might require a prohibitively large sample size. Rather than specifying a desirably narrow confidence interval width, the researcher could also try specifying the largest confidence interval width that would still provide useful information.

4.27 Sample Size Requirement for Desired Power

The sample size required to perform a paired-samples t-test with a specified α value and desired power is approximately

$$n = 2\tilde{\sigma}^2(1-\tilde{\rho}_{12})(z_{\alpha/2} + z_\beta)^2/(\tilde{\mu}_1 - \tilde{\mu}_2)^2 + \frac{z_{\alpha/2}^2}{2} \qquad (4.32)$$

where $\tilde{\rho}_{12}$ is a planning value of the Pearson correlation between the two measurements, $\tilde{\sigma}^2$ is a planning value of the average within-condition variance, and $\tilde{\mu}_1 - \tilde{\mu}_2$ is a planning value of the effect size. Note that the sample size requirement is larger for smaller values of $\tilde{\rho}_{12}$, smaller effect sizes, and greater desired power. Equation 4.32 can be computed using the `size.test.mean.ps` function in the `statpsych` package. In applications where the sample size is predetermined, the power.mean.ps function in the `statpsych` package will estimate the power of the test for the given sample size, α value, and effect size.

Multiplying the result from Equation 4.32 by 1.05 approximates the sample size requirement for the Wilcoxon signed rank test when the difference scores are approximately normal. If the difference scores are highly leptokurtic, this approximation gives a conservatively large sample size requirement.

The sample size required to perform a directional two-sided test for a linear contrast of population means in a within-subjects design with a specified α value and desired power is approximately

$$n = \tilde{\sigma}^2(\sum_{j=1}^{m} q_j^2)(1-\tilde{\rho})(z_{\alpha/2}+z_{\beta})^2/(\sum_{j=1}^{m} q_j\tilde{\mu}_j)^2 + \frac{z_{\alpha/2}^2}{2} \quad (4.33)$$

where $\tilde{\sigma}^2$ is a planning value for the average variance of the a measurements, $\tilde{\rho}$ is a planning value for the average correlation among all pairs of measurements, and $\sum_{j=1}^{m} q_j\tilde{\mu}_j$ is a planning value of the effect size. This formula assumes $\sum_{j=1}^{m} q_j = 0$. A Bonferroni adjustment to α in the critical z-value can be used when two or more simultaneous hypothesis tests will be performed.

In applications where $\sum_{j=1}^{m} q_j\tilde{\mu}_j$ or $\tilde{\sigma}^2$ is difficult to specify, Equation 4.33 can be expressed more simply in terms of a planning value for φ, as shown below

$$n = (\sum_{j=1}^{m} q_j^2)(1-\tilde{\rho})(z_{\alpha/2}+z_{\beta})^2/\tilde{\varphi}^2 + \frac{z_{\alpha/2}^2}{2} \quad (4.34)$$

which specializes to

$$n = 2(1-\tilde{\rho}_{12})(z_{\alpha/2}+z_{\beta})^2/\tilde{\delta}^2 + \frac{z_{\alpha/2}^2}{2} \quad (4.35)$$

for a paired-samples design or pairwise comparisons.

Equation 4.33 can be computed using the `size.test.lc.mean.ws` function in the `statpsych` package. Equation 4.34 can also be computed using the `size.test.lc.mean.ws` function by setting the variance planning value to 1 and the effect size to $\tilde{\varphi}$. For a paired-samples t-test, SPSS can compute the required sample size for the desired power. The strategies for specifying an effect size in section 1.29 can also be used to specify the effect size in Equations 4.32 and 4.33.

Example 4.17. A researcher is planning a 2 × 2 within-subjects facial recognition experiment and wants to reject the null hypothesis of a zero two-way interaction effect with power of .95 at α = .05. After conducting a pilot study and reviewing previous research, it was decided to set $\tilde{\sigma}^2$ = 15 and $\tilde{\rho}$ = 0.8. The expected size of the interaction contrast is 3.0. The required sample size is approximately $n = 15(4)(1 - 0.8)(1.96 + 1.65)^2/3.0^2 + 1.92 = 19.3 \approx 20$.

The sample size requirement to perform an equivalence test for the difference in two population means in a paired-samples design with a specified α value and desired power is approximately

$$n = 2\tilde{\sigma}^2(1 - \tilde{\rho})\left(z_{\alpha} + z_{\beta/2}\right)^2/(h - |\tilde{\mu}_1 - \tilde{\mu}_2|)^2 + \frac{z_{\alpha}^2}{2} \tag{4.36}$$

where $\tilde{\mu}_1 - \tilde{\mu}_2$ is the expected effect size which must be smaller than h where the interval $-h$ to h is the region of practical equivalence. Equivalence tests usually require large sample sizes. The `size.equiv.mean.ps` function in the `statpsych` package computes Equation 4.36.

Example 4.18. A researcher wants to show that two prototype navigation programs have similar usability mean ratings. A sample of participants will use both programs for 20 days and then rate each program on a 1 to 30 scale. The researcher believes that a 2-point difference in mean ratings is small and unimportant. The required sample size to perform an equivalence test with power of .9, α = .10, an expected effect size of 0.5, a variance planning value of 5, and a correlation planning value of .75 is approximately
$n = 2(25)(1 - .75)(1.28 + 1.65)^2/(2 - .5)^2 + 0.82 = 48.5 \approx 49$.

The sample size requirement to perform a superiority test for the difference in two population means in a paired-samples design with a specified α value and desired power is approximately

$$n = 2\tilde{\sigma}^2(1 - \tilde{\rho})\left(z_{\alpha/2} + z_{\beta}\right)^2/(\tilde{\mu}_1 - \tilde{\mu}_2 - h)^2 + \frac{z_{\alpha/2}^2}{2} \tag{4.37}$$

where $\tilde{\mu}_1 - \tilde{\mu}_2$ is the expected effect size which must be greater than h. Equation 4.37 can also be used for a noninferiorty test by replacing h with $-h$ and specifying a value for $\tilde{\mu}_1 - \tilde{\mu}_2$ that is greater than $-h$. Compared to Equation 4.32, the sample size given by Equation 4.37 will be larger for a superiority test and smaller for a

noninferiority test. The `size.supinf.mean.ps` function in the `statpsych` package computes Equation 4.37.

Equation 1.17 can be used to approximate the sample size required for a pared-samples sign test. In a paired-samples experiment, the effect size (π) is the proportion of individuals in the study population who would score higher under Treatment 1 than Treatment 2. The `size.test.sign.ps` function in the `statpsych` package will compute Equation 1.17 for a paired-samples design.

Example 4.19. A researcher is planning a paired-sample experiment and expects the proportion of individuals in the population who would score higher under Treatment 1 than Treatment 2 to be about .7. Using the `size.test.sign.ps` function with power = .80, α = .05, and a π planning value of .7, the required sample size is 52

4.28 Data Visualization

Estimated means or medians for each level of a within-subjects factor can be presented graphically using a bar chart or error bar chart with 95% confidence interval lines. The data must be in long format to produce this type of bar chart in SPSS. A bar chart of sample means with 95% confidence interval lines for a one-factor within-subject study is shown below.

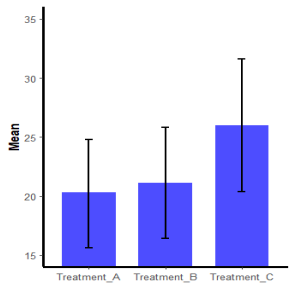


Figure 4.1 Bar chart for a within-subjects experiment

The confidence interval line at treatment *j* describes the range of plausible values for the population mean assuming everyone in the population had received treatment *j*. A bar chart of medians with confidence intervals lines has the same interpretation in terms of population medians.

In a within-subjects design, the confidence intervals for individual means can be substantially wider than the confidence intervals for mean differences, and it is informative to examine a bar chart for pairwise mean differences as shown below.

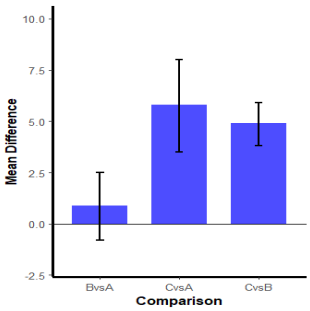


Figure 4.2 Bar chart for all pairs of mean differences

The sample means or medians for two-factor within-subject or mixed designs can be presented graphically using a clustered bar chart with 95% confidence intervals. Consider a 2 × 2 within-subjects study where participants tried to correctly identify a vowel or a consonant heard under low and high white noise conditions. The response variable is the number of errors in 80 trials. A clustered bar chart for the mean number of errors is shown below.

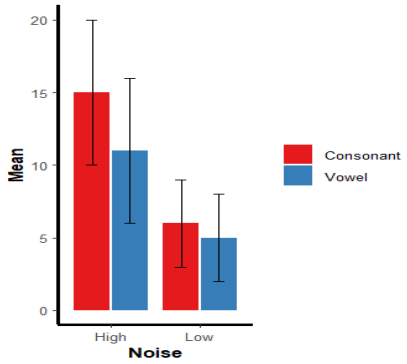


Figure 4.3 Clustered bar chart for a 2 × 2 within-subjects experiment

In this example, the comparison of vowels with consonants is more interesting than the comparison of low and high noise, and so the vowel and consonant conditions are displayed within the noise clusters. A second type of bar chart for this study could have two bars; one bar for the difference in consonants and vowels at high noise, and a second bar for the difference in consonants and vowels at low noise. The error bar charts one-factor and two factor designs described in section 3.27 are also useful in within-subjects and mixed designs.

In a two-factor mixed design, the between-subject factor is often used to define the cluster of bars and the within-subject factor defines the bars within each cluster. If the A × B interaction effect is small, a separate bar chart for each factor is an alternative to a clustered bar chart.

Dot plots and a side-by-side box plot can also be used in one-factor and factorial within-subjects designs to assist in the data screening process. The different levels of the within-subjects factor are treated like different groups in a side-by-side box plot. A dot plot could be examined for the y scores at each level of the within-subjects factor.

A *scatterplot* can be used for data screening in within-subjects designs to detect *bivariate outliers*. A bivariate outlier is a participant who has an unusual *combination* of responses at any two levels of a within-subjects factor. An example of a scatterplot for a paired-samples experiment is shown below, and it appears that the participant in the upper left corner of the plot is a bivariate outlier. Note that this outlier would have gone unnoticed in a box plot of the Treatment 1 scores or a box plot of the Treatment 2 scores.

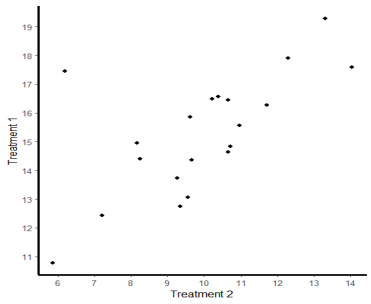


Figure 4.4 Scatterplot of paired scores in a paired-samples design

Key Terms

within-subjects experiment
within-subjects factor
carryover effect
linear contrast score
one-way within-subjects ANOVA
pretest-posttest designs
two-factor within-subjects experiment
two-way within-subjects ANOVA
mixed two-factor design (split plot design)
two-way mixed ANOVA
completely counterbalanced design
balanced Latin square counterbalancing
symmetric carryover effects
asymmetric carryover effects
compound symmetry assumption
measurement error
reliability coefficient
interrater reliability
alternate form reliability
test-retest reliability
internal consistency reliability
coefficient alpha (Cronbach's alpha)
Spearman-Brown formulas
listwise deletion
pairwise deletion
Wilcoxon signed rank test
Friedman test
Scatterplot

Concept Questions

1. How is a within-subjects design different from a between-subjects design?

2. Explain in words how a confidence interval for $\mu_1 - \mu_2$ in a paired-samples design is computed.

3. How does the correlation between measurements in within-subjects designs affect the sample size requirement for a confidence interval of $\mu_1 - \mu_2$?

4. What are the assumptions of the one-way within-subjects ANOVA hypothesis test for equal population means?

5. What are the assumptions for a confidence interval for a linear contrast of population means in a within-subjects design?

6. Why is a confidence interval for a linear contrast of means preferred to the test of equal population means in within-subjects designs?

7. How could a confidence interval for the AB interaction effect in a 2 × 2 mixed design be computed?

8. How can a confidence interval for the main effect of the between-subjects factor in a 2 × 2 mixed design be computed?

9. For a 2 × 2 within-subjects experiment, specify the linear contrast scores to estimate or test the following effects in terms of the four scores per participant (y_{11}, y_{12}, y_{21}, and y_{22}). The first subscript specifies the levels of Factor A and the second subscript specifies the levels of Factor B.

a) A × B interaction
b) main effect of A
c) main effect of B
d) simple main effect of A at b_1
e) simple main effect of A at b_2
f) simple main effect of B at a_1
g) simple main effect of B at a_2

10. What is one way to control for carryover effects?

11. What hypothesis tests can be performed using a confidence interval for $\mu_1 - \mu_2$ in a within-subjects experiment?

12. Why is a pretest-posttest design more ethical than a two-group experiment with a control group when the treatment is expected to be beneficial?

13. What are some important uses of a mixed design?

14. How does missing data affect the one-way within-subjects ANOVA vs. pairwise comparisons?

15. What distribution-free methods could be used to analyze data from a one-way within-subjects design? When would these methods be preferred to a one-way within-subjects ANOVA or pair-wise comparisons using paired-samples *t*-tests?

16. What is the effect of measurement error on the width of confidence intervals and the power of tests in between-subjects and within-subjects designs?

17. What is the effect of measurement error on estimates of δ, φ, and η^2?

18. The blood pressure measurement devices used in most behavioral labs are not highly reliable. Suppose the reliability of a single blood pressure reading is .81. What is the reliability of the average of two blood pressure measurements?

19. Fifty UCSC students were randomly selected from the entering class of about 4,500 freshman and given a public speaking examination (scored 0 to 100) in the Fall quarter of their first year (time 1). These 50 students were given the same examination during the Fall quarter of their third year (time 2). A 95% confidence interval for $\mu_1 - \mu_2$ was [-8.9, -3.7]. Interpret this result.

20. Explain why each of the following interpretations of [-8.9, -3.7] in the above example is incorrect or inadequate.

a) We are 95% confident that the mean public speaking score of the 50 students increased 3.7 to 8.9 points from their first year to their third year.

b) We are 95% confident that a typical UCSC student will have a public speaking score that will improve 3.7 to 8.9 points from their first year to their third year.

c) We are 95% confident that the mean public speaking score for all 4,500 freshman drops 3.7 to 8.9 points in their third year.

d) We are 95% confident that the difference in population means is between -8.9 and -3.7.

Data Analysis Problems

4-1. Eight 3rd year UCSC psychology students were randomly selected from the psychology department's student database which contains the names of about 600 3rd year psychology majors. Each student was asked to rate on a 1 to 50 scale their degree of satisfaction with their non-psychology courses, their required psychology courses, and their elective psychology courses. The ratings are given below.

Student	non-psych	req-psych	elect-psych
1	15	20	25
2	17	16	20
3	30	32	39
4	22	23	29
5	14	14	18
6	19	19	22
7	19	20	25
8	26	25	30

a) Describe the study population.

b) Describe the population means (μ_1, μ_2, and μ_3) in the context of this study.

c) Use SPSS to compute the multivariate test of H_0: $\mu_1 = \mu_2 = \mu_3$. Report *F*, degrees of freedom, and *p*-value in APA style.

d) Use SPSS or R to compute paired-samples *t*-tests for all three pairs of mean differences. Report *t*, degrees of freedom, and *p*-value in APA style for each test.

e) Use SPSS or R to compute Bonferroni 95% confidence interval for all three pairs of mean differences and interpret the results.

f) Use the `ci.stdmean.ps` function in the `statpsych` package to compute Bonferroni 95% confidence interval for all three pairs of standardized mean differences and interpret the results.

g) If the study is to be replicated at another university, how many 3rd year psychology students would be needed to obtain a 95% confidence interval for a difference of any two population means that has a width of about 1.0? Use the largest variance and the smallest correlation from this study as planning values.

4-2. Previous research has shown that scores on a driving simulator are lower when the driver uses a cell phone. A new study examined the effects of traffic conditions and the effects of sending or receiving a call on driving performance.

Twelve participants were randomly sampled from a volunteer pool of about 2,500 undergraduate students and tested in a driving simulator under four different conditions. Two of the conditions simulated light traffic, and two conditions simulated heavy traffic. Under each traffic condition, each participant was required to send a call and also answer a call on their own cell phone. The 12 participants were randomly assigned to four balanced Latin square order conditions. Assume the practice, fatigue, and carryover effects are negligible so that the between-subject order factor can be ignored. Driving performance scores (higher scores represent better performance) were obtained for 2-minute periods following the sending and receiving of calls in each of the two traffic conditions. The driving performance scores are shown below.

		Light Traffic		Heavy Traffic	
Student	Order	Receive	Send	Receive	Send
1	1	21	20	21	17
2	1	39	36	36	33
3	1	32	33	30	28
4	2	29	27	27	27
5	2	27	28	28	27
6	2	17	14	15	16
7	3	27	30	27	26
8	3	21	20	18	20
9	3	28	27	29	25
10	4	17	15	16	15
11	4	12	11	11	13
12	4	27	22	22	22

a) Describe the study population.

b) Describe the population means (μ_1, μ_2, μ_3, and μ_4) in the context of this study.

c) Use SPSS or R to compute a test of the two-way within-subject interaction using a linear contrast score. Report t, degrees of freedom, and p-value in APA style.

d) Use SPSS or R to compute Bonferroni 95% confidence intervals for the two main effects. Interpret the results.

4-3. Eighteen participants were randomly selected from a university research participant pool consisting of about 3,800 undergraduate students. The 18 participants were randomly divided into three groups of equal size. In the first two groups, participants were briefly presented with 20 3-letter and 20 5-letter strings in random order and then asked if the string included some particular letter. In group 1, the letter strings were 3-letter or 5-letter words (e.g., cat, great).

In group 2, the strings were 3-letter or 5-letter non-words (e.g., tca, aetgr). In group 3, participants were briefly presented with 20 3-digit and 20 5-digit numbers in random order and then asked if the number included some particular digit. The researcher wants to show that it is easier to determine if a particular letter is contained in a word than in a non-word. The researcher also believes that strings of 3 and 5 numbers are encoded in a manner similar to words so that a 3-digit number such as 527 would be encoded as a single entity "five hundred and twenty-seven" rather than "five-two-seven". Each participant received a score from 0 to 20, representing the number of correct responses. The scores are shown below.

Participant:	1	2	3	4	5	6	
3 letter words	**19**	**20**	**20**	**20**	**20**	**19**	**Group 1 (words)**
5 letter words	**19**	**18**	**19**	**20**	**17**	**16**	

Participant:	7	8	9	10	11	12	
3 letter non-words	**19**	**16**	**16**	**14**	**16**	**18**	**Group 2 (non-words)**
5 letter non-words	**16**	**10**	**12**	**9**	**13**	**15**	

Participant:	13	14	15	16	17	18	
3 digit number	**20**	**17**	**19**	**20**	**19**	**20**	**Group 3 (numbers)**
5 digit number	**20**	**16**	**17**	**19**	**18**	**19**	

a) Describe the study population.

b) Describe the population means ($\mu_1, \ \ldots \ , \mu_6$) in the context of this study.

c) Use SPSS or R to test the string length by string type interaction (this can be done using a two-way mixed ANOVA or a one-way ANOVA on the difference scores). Report F, degrees of freedom, and p-value in APA style for the interaction effect. What effects should be examined next?

d) Use SPSS or R to compute Bonferroni 95% confidence intervals for the simple main effects of string length at words, non-words, and numbers. Interpret the results.

Suggested Readings

Hollander, M. & Wolfe, D. A. (1999). *Nonparametric statistical methods (2nd ed.).* Wiley. **Chapter 3**

Howell, D. C. (2013). *Statistical methods for psychology (8th ed.)* Thompson & Wadsworth. **Section 7.4 and Chapter 14**

Kepple, G. & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook (4th ed).* Prentice Hall. **Part IV and Chapter 23**

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2024). *Designing and analyzing data. A model comparison perspective (3rd ed)*. Lawrence Erlbaum. **Chapters 11, 12, 13, and 14**

Snedecor, G. W. & Cochran, W. G. (1989). *Statistical methods (8th ed.).* Iowa State Press. **Sections 6.1, 6.2, 6.3, and 6.4**

Appendix A. Tables

Table 1a Two-sided critical z-values ($z_{\alpha/2}$)

$1 - \alpha$				
.80	.90	.95	.99	.999
1.28	**1.65**	**1.96**	**2.58**	**3.29**

Table 1b One-sided critical z-values (z_{α})

$1 - \alpha$				
.80	.90	.95	.99	.999
0.84	**1.28**	**1.65**	**2.33**	**3.09**

R Functions

Use **qnorm(1 - α/2)** for 2-sided critical z-value.

Example: α = .005

```
qnorm(1 - .005/2)
2.807034
```

Use **qnorm(1 - α)** for 1-sided critical z-value.

Example: β = .25

```
qnorm(1 - .25)
0.6744898
```

Use **2*(1 - pnorm(abs(z)))** to compute 2-sided p-value for z statistic.

Example: z = 2.32

```
2*(1 - pnorm(2.32))
0.02034088
```

Table 2 Two-sided critical *t*-values ($t_{\alpha/2;df}$)

	1 − α			
df	.90	.95	.99	.999
1	**6.31**	**12.71**	**63.66**	**636.62**
2	**2.92**	**4.30**	**9.93**	**31.60**
3	**2.35**	**3.18**	**5.84**	**12.92**
4	**2.13**	**2.78**	**4.60**	**8.61**
5	**2.02**	**2.57**	**4.03**	**6.86**
6	**1.94**	**2.45**	**3.71**	**5.96**
7	**1.90**	**2.37**	**3.50**	**5.41**
8	**1.86**	**2.31**	**3.36**	**5.04**
9	**1.83**	**2.26**	**3.25**	**4.78**
10	**1.81**	**2.23**	**3.17**	**4.59**
11	**1.80**	**2.20**	**3.11**	**4.48**
12	**1.78**	**2.18**	**3.06**	**4.32**
13	**1.77**	**2.16**	**3.01**	**4.22**
14	**1.76**	**2.15**	**2.98**	**4.14**
15	**1.75**	**2.13**	**2.95**	**4.07**
16	**1.75**	**2.12**	**2.92**	**4.02**
17	**1.74**	**2.11**	**2.90**	**3.97**
18	**1.73**	**2.10**	**2.88**	**3.92**
19	**1.73**	**2.09**	**2.86**	**3.88**
20	**1.73**	**2.09**	**2.85**	**3.85**
25	**1.71**	**2.06**	**2.79**	**3.73**
30	**1.70**	**2.04**	**2.75**	**3.65**
35	**1.69**	**2.03**	**2.72**	**3.59**
40	**1.68**	**2.02**	**2.70**	**3.55**
60	**1.67**	**2.00**	**2.66**	**3.46**
∞	**1.65**	**1.96**	**2.58**	**3.29**

R Functions

Use **`qt(1 - α/2, df)`** for 2-sided critical *t*-value.

Example: α = .05 and *df* = 22

```
qt(1 - .05/2, 22)
2.073873
```

Use **`2*(1 - pt(abs(t), df))`** to compute 2-sided *p*-value for *t* statistic.

Example: *t* = -2.43 and *df* = 9

```
2*(1 - pt(2.43, 9))
0.03798256
```

Table 3 Two-sided Bonferroni critical z-values ($z_{\alpha*/2}$)

Simultaneous Confidence Level	Number of Confidence Intervals 2	3	4	5	6	7	8	9	10
.90	**1.96**	**2.13**	**2.24**	**2.31**	**2.40**	**2.45**	**2.50**	**2.54**	**2.58**
.95	**2.24**	**2.39**	**2.50**	**2.58**	**2.64**	**2.69**	**2.74**	**2.77**	**2.81**
.99	**2.81**	**2.94**	**3.02**	**3.09**	**3.15**	**3.19**	**3.23**	**3.26**	**3.39**

Note: These critical values can be used to obtain simultaneous confidence intervals for standardized mean differences, Mann-Whitney parameters, MADs, or ratios of MADs.

R Function

Use **qnorm(1 - α/(2 * g))** for 2-sided critical z-value and g simultaneous confidence intervals.

Example: α = .05 and g = 4

```
qnorm(1 - .05/(2*4))
2.497705
```

Two-sided Bonferroni critical t-values ($t_{\alpha*/2;df}$)

R Function

Use **qt(1 - α/(2 * g), df)** for 2-sided critical t-value and g simultaneous confidence intervals.

Example: α = .05, g = 4, and df = 15

```
qt(1 - .05/(2*4), 15)
2.836627
```

Table 4 Critical *F* values (α = .05, upper tail)

	df_A			
df_E	1	2	3	4
2	**18.51**	**19.00**	**19.16**	**19.25**
3	**10.13**	**9.55**	**9.28**	**9.12**
4	**7.71**	**6.94**	**6.59**	**6.39**
5	**6.61**	**5.79**	**5.41**	**5.19**
6	**5.99**	**5.14**	**4.76**	**4.53**
7	**5.59**	**4.74**	**4.35**	**4.12**
8	**5.32**	**4.46**	**4.07**	**3.84**
9	**5.12**	**4.26**	**3.86**	**3.63**
10	**4.96**	**4.10**	**3.71**	**3.48**
11	**4.84**	**3.98**	**3.59**	**3.36**
12	**4.75**	**3.89**	**3.49**	**3.26**
13	**4.67**	**3.81**	**3.41**	**3.18**
14	**4.60**	**3.74**	**3.34**	**3.11**
15	**4.54**	**3.68**	**3.29**	**3.06**
16	**4.49**	**3.63**	**3.24**	**3.01**
17	**4.45**	**3.59**	**3.20**	**2.96**
18	**4.41**	**3.55**	**3.16**	**2.93**
19	**4.38**	**3.52**	**3.13**	**2.90**
20	**4.35**	**3.49**	**3.10**	**2.87**
∞	**3.84**	**3.00**	**2.61**	**2.37**

R Functions

Use **`qf(1 - α, df1, df2)`** for critical *F* value.

Example: α = .01, *df1* = 2, and *df2* = 27

```
qf(1 - .01, 2, 27)
5.488118
```

Use **`1 - pf(F, df1, df2)`** to compute *p*-value for *F* statistic.

Example: *F* = 4.74, *df1* = 2, and *df2* = 10

```
1 - pf(4.74, 2, 10)
0.03564965
```

Table 5 Critical chi-square values

	α (lower tail)			α (upper tail)		
df	.05	.10	.20	.10	.05	.01
1	.004	0.02	0.06	2.71	3.84	6.63
2	0.10	0.21	0.45	4.61	5.99	9.21
3	0.35	0.58	1.00	6.25	7.81	11.34
4	0.71	1.06	1.65	7.78	9.49	13.28
5	1.14	1.61	2.34	9.24	11.07	15.09
6	1.63	2.20	3.07	10.64	12.59	16.81
7	2.17	2.83	3.82	12.02	14.07	18.48
8	2.73	3.49	4.59	13.36	15.51	20.09
9	3.32	4.17	5.38	14.68	16.92	21.67
10	3.94	4.86	6.17	15.99	18.31	23.21
11	4.57	5.58	6.99	17.28	19.68	24.72
12	5.23	6.30	7.81	18.55	21.03	26.22
13	5.89	7.04	8.63	19.81	22.36	27.69
14	6.57	7.79	9.47	21.06	23.68	29.14
15	7.26	8.55	10.31	22.31	25.00	30.58
16	7.96	9.31	11.15	23.54	26.30	32.00
17	8.67	10.08	12.00	24.77	27.59	33.41
18	9.39	10.85	12.86	25.99	28.87	34.81
19	10.11	11.65	13.72	27.20	30.14	36.19
20	10.85	12.44	14.57	28.41	31.41	37.57

R Functions

Use **`qchisq(1 - α, df)`** for critical chi-square value.

Example: α = .05 and df = 7.

```
qchisq(1 - .05, 7)
14.06714
```

Use **`1 - pchisq(c, df)`** to compute *p*-value for chi-square statistic.

Example: chi-squared = 9.14 and df = 4.

```
1 - pchisq(9.14, 4)
0.05769384
```

Appendix B. Glossary

adjusted eta-squared – a correction to an eta-squared estimate that reduces its positive bias

alternate form reliability – the correlation between two alternative forms of a test or questionnaire that each assess the same attribute

asymmetric carryover effect – when the carryover effect from Treatment 1 to Treatment 2 (for example) is not the same as the carryover effect from treatment 2 to 1 in a within-subjects design

balanced Latin square counterbalancing – a partial counterbalanced within-subjects design that controls for practice effects, fatigue effects, and certain types of carryover effects

bar chart – a graph that illustrates the parameter estimates (e.g., sample means) in two or more groups using bars with lengths that represents the values of the estimates

between-subjects treatment factor – an independent variable in a between-subjects experiment where participants are randomly assigned to the levels of the independent variable

Bonferroni adjustment – used to obtain simultaneous tests or confidence intervals by dividing α by the number of tests or confidence intervals to be examined

box plot – a graph that shows the 25^{th}, 50^{th} (median), and 75^{th} percentiles and possible outliers

carryover effect – when participants are exposed sequentially to more than one treatment, the effect of one treatment persists during other treatments

central limit theorem – a theorem stating that the shape of a sampling distribution of a mean has an approximate normal distribution if the sample size is sufficiently large regardless of the shape of the distribution of quantitative scores in the population

classification factor – a type of factor in which participants are classified into two or more groups according to some participant characteristic such as conservative/liberal or male/female

clustered bar chart – a graph that illustrates the means in a two-factor design using bars with lengths that represents the values of the means

coefficient alpha (Cronbach's alpha) – a measure of the reliability of a sum or average of two or more measurements

coefficient of skewness – a measure of the asymmetry of a set of quantitative scores; the coefficient is equal to zero if the scores have a symmetric distribution

coefficient of kurtosis – a measure of the peakedness and tail thickness in a distribution of quantitative scores

Cohen's d – a mean difference divided by the square root of a pooled variance estimate

completely counterbalanced design – a within-subjects design where subsets of participants receive the within-subject treatments in one of every possible order

confidence interval – a range of values that will include the population parameter, or function of population parameters, with a specified level of confidence

confidence interval width – the upper limit minus the lower limit of a confidence interval

confidence level – a subjective probability that a computed confidence interval will contain the value of some unknown population parameter

confounding variable – a variable that has a causal effect on the response variable and the independent variable

contrast coefficient – the $v_1, v_2, \ldots, v_a$ values that define a linear contrast (e.g., $v_1\mu_1 + v_2\mu_2 + \cdots + v_a\mu_a$) where $\sum_{j=1}^{a} v_j = 0$

control group – a group of participants who do not receive any treatment

data transformation – a nonlinear transformation of scores, such as $ln(y)$, $1/y$, $\sqrt{y}$, that can reduce the degree of nonnormality in a set of scores

debriefing – a short interview with the participants immediately following completion of a study to ensure that the that participants are fully informed about, and not harmed in any way by, their experience in the study

dependent variable – a variable assumed to be predictable from by one or more independent variables (the dependent variable is also called a *response variable* or *outcome variable*)

differential nonrandom attrition – when the independent variable causes certain types of participants to drop out with different probabilities across the treatment conditions

directional error – accepting one alternative hypothesis when the other alternative hypothesis is true (e.g., accepting H_1: $\mu_1 > \mu_2$ when H_2: $\mu_1 < \mu_2$ is true)

directional two-sided hypothesis – a test of a null hypothesis with two directional alternative hypotheses

dot plot – a graph of the y scores of each participant

effect size – some examples of an effect size are: $\mu - h$, $\mu_1 - \mu_2$, $\sum_{j=1}^{a} v_j\mu_j$, standardized mean difference, standardize linear contrast of means, eta-squared, difference of medians

equivalence test – a test of one hypothesis that the difference between two parameters (e.g., means, medians) is small against another hypothesis that the difference is not small

error variance – in a multiple group design, the error variance is the within-group variance

eta-squared – a standardized measure of effect size in a one-way ANOVA that describes the proportion of variance in the response variable that is predictable by the independent variable

experimental design – a study where participants are randomly assigned into the levels of an independent variable

exploratory data analysis – examining multiple response variables and/or independent variables in search of "significant" results

external validity – the extent to which the results of a study apply to different types of participants and different settings

F ratio – a test statistic; in a one-way ANOVA the *F* ratio is equal to the between-group mean square divided by the within-group (error) mean square

factor – an independent variable with a predetermined number of levels

factorial design – a design with two or more factors

family-wise directional error rate – the probability of making one or more directional errors in a set of two or more directional two-sided hypothesis tests

F test – The use of an *F* statistic from an ANOVA table to test a null hypothesis that some effect is zero

Fisher protected test – the use of non-simultaneous tests of pairwise comparisons for a particular factor only if the omnibus null hypothesis for the factor has been rejected; recommended only for factors with three levels

Friedman test – a distribution-free alternative to the one-way within-subjects ANOVA

Games-Howell – simultaneous tests and confidence intervals for all pair-wise differences of populations means that does not assume equal population variances

generalized eta-squared – a standardized measure of effect size that describes the proportion of variance in the response variable that is predictable by the independent variable without removing the variability in the response variable that is due to any of the classification factors in the design

histogram – a graph of the number of quantitative scores that fall into specified intervals and illustrates the shape of the distribution of quantitative scores

Holm method – a method of testing multiple hypotheses that keeps the family-wise direction error rate at some specified value

homoscedasticity assumption – the assumption of approximately equal population variances across treatment conditions or subpopulations

independence assumption – the assumption that the response of each participant in the sample has no effect on the response of any other participant in the sample

independent-samples t-test – a test of H_0: $\mu_1 = \mu_2$ where μ_1 and μ_2 are unknown population means that are estimated from two different groups of participants.

independent variable – a variable that is assumed to predict the response variable (the independent variable is also called the *explanatory variable* or *predictor* variable)

informed consent – making all relevant information available before obtaining consent to participate in a study

interaction effect – the difference in effects of one factor across the levels of a second factor (e.g., in a 2 × 2 design, the interaction effect is $(\mu_{11} - \mu_{12}) - (\mu_{21} - \mu_{22})$)

internal consistency reliability – the reliability of a multi-item questionnaire that uses the individual item responses to estimate the reliability

internal validity – the extent to which a predictor variable can be assumed to have a causal effect on the response variable

interrater reliability – the correlation between two ratings of the same attribute

interval scale – a scale where a difference in scores correctly describes a difference in the attribute but a score of 0 does not represents a complete absence of the attribute

IRB (Institutional Review Board) – a committee of experts who evaluate the ethical issues in studies involving human subjects

linear contrast (of means) – a linear combination of means $v_1\mu_1 + v_2\mu_2 + \cdots + v_k\mu_k$ where the $v_1, v_2, \ldots, v_k$ values are contrast coefficients specified by the researcher

linear contrast score – a linear combination of scores $q_1y_1 + q_2y_2 + \cdots + q_ay_a$ in a within-subject design where the $q_1, q_2, \ldots, q_a$values are contrast coefficients specified by the researcher

listwise deletion – deleting a participant from the analysis in a within-subjects design if the participant has missing scores at any of the within-subject levels

main effect – the differences in mean values across the levels of one factor across the levels of all other factors in the design

Mann-Whitney test – a distribution-free alternative to the independent-samples *t*-test

matched-pairs design – a 2-level within-subjects design where each pair of similar participants is randomly assigned to two treatment conditions

mean absolute deviation from median (MAD) – a measure of variability that is equal to the average of the absolute deviations of the quantitative scores from the median

mean squared error – an estimate of the average within-group variance

measurement error – the difference between a person's true attribute score and some measurement of the attribute

minimal risk – a level of risk that is similar to that which would be incurred during a routine physical or psychological examination

minimally interesting effect size – the smallest value of an effect size that would still represent an interesting, useful, or important effect size

mixed two-factor design – a design with one between-subjects factor and one within-subjects factor

mixed two-way ANOVA – used to test three separate hypotheses in a mixed two-factor: equal population means across levels of Factor A, equal population means across levels of Factor B, and zero interaction between Factors A and B

monotonic transformation – a transformation that does not change the rank ordering of the scores

nondifferential nonrandom attrition – when a certain type participant is more likely to drop out of a study than other participants but the dropout rate is equal across treatment conditions

nonexperimental design – a study where participants are not randomly assigned to the levels of an independent variable

noninferiority test – a test of one hypothesis that Treatment 1 is not unacceptably worse than Treatment 2 against another hypothesis that Treatment 1 is unacceptably worse than Treatment 2

non-removable interaction – an interaction effect that cannot be made small or unimportant by transforming the response variable

normal (Gaussian) curve – a symmetric bell-shaped curve

normality assumption – the assumption that the scores have an approximate normal (Gaussian) distribution in the study population

null hypothesis – a statement that some population parameter, or some function of populations parameters (e.g., a difference or linear contrast) is equal to a specific numerical value

omnibus test – a test of a null hypothesis that specifies the quality of two or more parameters; tests of main effects and interaction effects for factors with three or more levels are examples of omnibus tests

one-sample t-test – a test of H_0: $\mu = h$ where μ is an unknown population mean and h is a specified number; test is used to decide if $\mu > h$ or if $\mu < h$

one-sample treatment design – a study where one sample of participants is obtained from a study population and each participant receives the same treatment

one-way ANOVA – a test of H_0: $\mu_1 = \mu_2 = ... = \mu_a$ in a between-subjects design where μ_j is an unknown population mean

one-way random effects ANOVA – a test of H_0: $\mu_1 = \mu_2 = ... = \mu_M$ in a between-subjects design where μ_j is an unknown subpopulation mean and M is a large set of subpopulations

one-way within-subjects ANOVA – a test of H_0: $\mu_1 = \mu_2 = ... = \mu_a$ in a within-subjects design where μ_j is an unknown population mean

paired-samples t-test – a test of H_0: $\mu_1 = \mu_2$ where μ_1 and μ_2 are unknown population means that are estimated from a single group and is used to decide if $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$

pairwise comparison – a difference between two parameter values (e.g., $\mu_1 - \mu_2$, $\mu_1 - \mu_4$, etc.)

pairwise deletion – deleting a participant from a within-subject analysis that involves a pair of within-subject levels if the participant has a missing score on either of the two levels

pairwise interaction effect – an interaction effect at two levels of one factor and two levels of a second factor

pairwise main effect comparison – a difference in means for two levels of one factor after averaging over the levels of a second factor

pairwise simple main effects – a difference in means for two levels of one factor at one level of a second factor

partial eta squared – a measure of effect size in a factorial ANOVA that describes the proportion of response variable variance than can be predicted by one treatment factor after removing the variability in the response variable due to all other treatment factors in the design

planning value – an assumed value of a population parameter, obtained from previous research or expert opinion, that is used in sample size formulas

population – in psychological research, the population is usually some well-defined large group of people that the researcher wants to study

population parameter – a single number, such as a population mean, population variance, or population median, that is defined for all members of a specific population

population mean – the mean of the response variable for all members of the population

population median – a value that is greater than half the scores and less than half the scores in a population

power – the probability that a statistical test will lead to the rejection of the null hypothesis

practice effect – an improvement in posttest scores that is the result of taking a pretest

prediction interval – a range of values that will include the score for a single person

predictor variable – a variable that may account for some of the variability in the response variable (the predictor variable is also called the *explanatory variable* or *independent* variable)

preregistration – a time-stamped online public disclosure of all aspects of a planned study

pretest-posttest design – a within-subjects design where participants are measured on two or more occasions and receive a treatment sometime after the first occasion and before the last occasion

probability – a number in the range 0 to 1 that describes the likelihood of a specific event

p-value – a number that is a transformation of the test statistic and used to decide if a null hypothesis can be rejected (e.g., if the *p*-value is less than .05 then reject H_0)

quantitative score – a score that is measured on either an interval or a ratio scale

random attrition – a random loss of participants

random factor – a factor where the levels have been randomly selected from a large set of possible levels

random sample – a subset of the population selected in a way that every sample of given size has the same chance of being selected

random sampling assumption – the assumption that the sample used to compute a test or confidence interval is a random sample or assumed to be a random sample

randomized – random assignment of participants into the levels of an independent variable

ratio scale – a scale where a score of 0 represents a complete absence of the attribute and where differences and ratios of scores correctly describe differences and ratios of the attribute

region of practical equivalence – a range of values for a difference in population parameters, –*h* to *h*, where *h* is a value that represents a small or unimportant difference

reliability coefficient – an index defined on a range of 0 to 1 that describes the proportion of variance in the true scores that can be predicted by the observed scores

removable interaction – an interaction effect that becomes small or unimportant after the response variable has been transformed

response variable – a variable to be explained or predicted (also called a *dependent variable* or *outcome variable*)

sample mean –the sum of *n* quantitative scores divided by *n*

sample standard deviation – the square root of the sample variance

sample variance – the mean of *n* squared deviations from the sample mean

sampling distribution – a distribution of parameter estimates for all possible samples of a given size

scatterplot – a two-dimensional plot of paired observations that can be used to detect bivariate outliers in a within-subjects design

Scheffé method – provides simultaneous confidence and controls FWDER for all possible linear contrasts in a one-factor design and is useful in exploratory research

scientific misconduct – inappropriate or unethical behavior of the researcher

sign test – a distribution-free alternative to the one-sample *t*-test

simple main effect – the differences in mean values across levels of one factor at a specific level of the second factor in a two-factor design

simple-simple main effect – the differences in mean values across levels of one factor at a specific combination of levels of a second and third factor in a three-factor design

Spearman-Brown formulas – classic psychometric formulas that show the relation between the reliability of a single measurement and the reliability of a sum or average of multiple measurements

standard error – the standard deviation of a sampling distribution; a measure of the precision of a parameter estimate and used in the construction of confidence intervals

standardized linear contrast – a linear contrast of means divided by a standard deviation

stratified random sampling – taking a random sample from two or more subpopulations

study population – the population of units (e.g., people) from which the random sample was taken (also called the *sampling frame*)

superiority test – a test of one hypothesis that Treatment 1 is superior by an nontrivial amount to Treatment 2 against another hypothesis that Treatment 1 is not superior by a nontrivial amount to Treatment 2

two-way interaction effect – a measure of how the effect of one factor differs across the levels of a second factor at one level of a third factor

subpopulation – a subset of the study population

symmetric carryover effect – when the carryover effect from Treatment 1 to Treatment 2 (for example) is the same as the carryover effect from Treatment 2 to Treatment 1 in a within-subjects design

target population – a population of theoretical interest that subsumes the study population

test-retest reliability – the correlation between a specific test or questionnaire that was given to each participant on two different occasions

test statistic – a number computed from the sample that is used in a decision rule to reject or fail to reject a null hypothesis

three-factor experiment – an experiment with three independent variables

three-way ANOVA – used to test hypotheses regarding the main effect of Factor A, main effect of Factor B, main effect of Factor C, AB interaction, AC interaction, BC interaction, and ABC interaction; also used to obtain estimates of partial or generalized eta-squared coefficients

three-way interaction – a difference in simple two-way interactions across the levels of a third factor

treatment factor – a factor with treatment levels to which participants have been randomly assigned

Tukey-Kramer method – simultaneous tests and confidence intervals for all pair-wise differences of populations means (the classical version assumes equal population variances)

two-factor within-subjects design – a study with two within-subject factors and no between-subject factors

two-stage cluster sampling – taking random samples from a random sample of subpopulations

two-way ANOVA – used to test three separate hypotheses in a two-factor between-subjects design: equal population means across levels of Factor A, equal population means across levels of Factor B, and zero interaction between Factors A and B; also used to obtain estimates of partial eta-squared coefficients

two-way interaction effect – a measure of how the effect of one factor differs across the levels of a second factor

two-way ANOVA – used to test three separate hypotheses in a two-factor within-subjects design: equal population means across levels of Factor A, equal population means across levels of Factor B, and zero interaction between Factors A and B

unbiased estimate – an estimate where the mean of its sampling distribution is equal to the population parameter value

waitlist control group – a control group that is given a potentially beneficial treatment after they have been compared with the treated group

Welch F test – an alternative to the one-way ANOVA that does not assume equal variances

Wilcoxon signed rank test – a distribution-free alternative to the paired-samples *t*-test

within-subjects factor – a factor in which the same group of participants are measured at all levels of the factor

within-subjects experiment – an experiment where each participant receives all treatment conditions, usually in counterbalanced order (also called a *randomized block design*)

Zelen design – a more ethical design that allows participants who were randomly assigned to a one treatment condition to switch to a preferred treatment condition

2 × 2 *factorial experiment* – an experiment with two independent variables where each independent variable has two levels

Appendix C. Answers to Concept Questions

Chapter 1

1. Explain in words how to compute a sample variance.

Compute the sample mean, subtract each score from the sample mean and square the difference, sum the n squared differences and divided the sum by $n - 1$.

2. Increasing the sample size will have what effect on the width of the confidence interval?

The confidence interval width will tend to decrease

3. Increasing the level of confidence will have what effect on the width of the confidence interval?

The confidence interval width will increase

4. Increasing the sample size will have what effect on the value of the standard error?

The SE will tend to decrease

5. Explain in words how a confidence interval for μ can be used to test a directional two-sided hypothesis where H_0: $\mu = 100$.

If the lower limit is greater than 100 then reject the null hypothesis and accept the alternative hypothesis that $\mu > 100$. If the upper limit is less than 100 then reject the null hypothesis and accept the alternative hypothesis that $\mu < 100$.

6. Increasing the sample size will have what effect on the power of a test?

The power will increase

7. Explain in words how to compute a 95% confidence interval for a population mean after you have computed the sample mean and its standard error in a sample of $n = 20$.

Find the critical t-value for alpha = .05 and $df = 19$. Multiply the critical t-value by the standard error. Add this product to the sample mean to get the upper limit and subtract this product to get the lower limit.

8. When planning a future study to estimate μ, what is the effect of decreasing the desired confidence interval width on the sample size requirement?

A larger sample size requirement

9. When planning a future study to estimate μ, what is the effect of increasing the desired level of confidence on the sample size requirement?

A larger sample size requirement

10. When planning a future study to estimate μ, what is the effect of using $\tilde{\sigma}^2 = 50$ rather than $\tilde{\sigma}^2 = 75$ on the sample size requirement?

A smaller sample size requirement

11. When planning a future study to test H_0: $\mu = h$, how does the desired power affect the sample size requirement?

Greater desired power requires a larger sample size

12. When planning a future study to test H_0: $\mu = h$, how does the α value affect the sample size requirement?

A smaller α value requires a larger sample size

13. Why are narrow confidence intervals desirable?

Narrower confidence intervals are more informative because they provide a more precise description of the population parameter

14. Why are confidence interval results more informative than a directional two-sided hypothesis test?

Because the confidence interval results provide information about the value of the population parameter while the directional two-sided test can only tell us if the parameter is less than some value or greater than some value.

15. Describe the standard error in the context of a sampling distribution.

The standard error is equal to the standard deviation of the sampling distribution.

16. Why is the 95% level of confidence a popular choice?

It provides a good compromise between confidence and precision.

17. What are the assumptions of a confidence interval or test for μ, and what are the effects of violating those assumptions?

Random sample, independence among participant, and approximate normality of the response variable. Violating the random sample and independence assumptions can invalidate the results. Violating the normality assumption is usually not a problem unless the sample size is small.

18. What are some ways to obtain a planning value for σ?

Expert opinion, prior research, pilot study, or if the response variable range is known (max – min)/4

19. Explain why a confidence interval might be uninterpretable if a nonrandom sample is used.

Unless the random can be assumed to be random sample from some describable population, the confidence interval and hypothesis test results are uninterpretable.

20. Suppose a 95% confidence interval for μ was [1.5, 4.8]. In a test of the null hypothesis H_0: $\mu = h$, for which of the following values of b would H_0 be rejected: 1.20, 3.30, 4.1, or 5.0?

1.20 and 5.0 because these values are outside the confidence interval range

21. How does the size of the study population effect the sample size requirement when n/N is small?

Virtually no effect

22. Why are data transformations sometimes used?

To reduce nonnormality of the response variable.

23. Is it appropriate to declare a null hypothesis such as H_0: $\mu = h$ to be true if the results are "nonsignificant"? Why?

No, because a failure to reject this type of null hypothesis (i.e., a "nonsignificant" result) does not imply that the null hypothesis is true. Also, this type of null hypothesis is almost never true because it is virtually impossible for μ to exactly equal the value of h.

24. When testing a null hypothesis such as H_0: $\mu = h$, what can be said about a "significant" result?

This only indicates that $\mu \neq h$

25. What is the relation between the size of the p-value and the sample size?

Increasing the sample size tends to decrease the p-value.

26. When would a confidence interval for a population median be preferred to a confidence interval for a population mean?

When the response variable is highly skewed

27. Why is it important to assess the variability of the response variable and not just the mean or median of the response variable?

The mean and median describe the center of a distribution of scores and can be misinterpreted as a description of how every member of the population would score. A measure of variability provides additional information about individual differences among members of the population.

28. What are the implications of the central limit theorem in terms of how a confidence interval for μ will perform?

The confidence interval for μ requires the sampling distribution to be approximate normal, and the central limit theorem guarantees that the sampling distribution will be approximate normal if the sample size is sufficiently large.

29. About 1.8 million people visit the Monterey Bay Aquarium each year. Exit interviews were conducted in 2019 for a random sample of 2,417 visitors. One question asked respondents to rate the quality of the educational experience on a 1 to 10 scale. The 95% confidence interval for μ was [8.78, 8.89]. Interpret this result.

We are 95% confident that the mean educational experience rating of the 1.8 million visitors is between 8.78 and 8.89.

30. Explain why each of the following interpretations of [8.78, 8.89] in the above example is incorrect or inadequate.

a) **We are 95% confident that the 1.8 million visitor ratings are between 8.78 and 8.89.**
b) **We are 95% confident that a visitor's rating will be between 8.78 and 8.89.**
c) **We are 95% confident that the mean rating of the 2,417 visitors is between 8.78 and 8.89.**
d) **We are 95% confident that μ is between 8.78 and 8.89.**
e) **We are 95% confident that 95% of the scores are between 8.78 and 8.89.**

a) The 1.8 million ratings would be between 1 and 10. The confidence interval describes the plausible range of the *mean* of the 1.8 ratings.

b) This interpretation is correct for a prediction interval but not for a confidence interval.

c) The confidence interval describes the mean of the 1.8 million ratings and not the mean of the 2,417 ratings.

d) The interpretation is inadequate because μ needs to be defined for the reader.

e) 95% of the scores would be in a much wider range. The confidence interval describes the plausible range of the *mean* of the 1.8 ratings.

Chapter 2

1. What are the three conditions that must be satisfied to show that the independent variable is causally related to the response variable?

The two variables must be related, there exists no other variable that is related to both the independent and response variable, and variation of the independent variable must occur prior any observed change in the response variable.

2. Explain how the results of an experiment can provide evidence that the independent variable has a causal effect on the response variable.

A two-group experiment can show if there is a relation between the independent variable and the response variable; since participants are randomly assigned to groups, there can be no other variable that is related to the independent variable; and participants are first exposed to different levels of the independent variable and then later their responses are recorded.

3. **How can a confidence interval for $\mu_1 - \mu_2$ be used to test H_0: $\mu_1 - \mu_2 = 0$ and then select H_1: $\mu_1 - \mu_2 > 0$ or H_2: $\mu_1 - \mu_2 < 0$?**

If the lower limit for $\mu_1 - \mu_2$ is greater than 0, then reject the null hypothesis and accept H_1: $\mu_1 - \mu_2 > 0$. If the upper limit for $\mu_1 - \mu_2$ is less than 0, then reject the null hypothesis and accept H_2: $\mu_1 - \mu_2 < 0$.

4. How can a confidence interval for δ be used to test H_0: $|\delta| \leq 0.1$ against H_1: $|\delta| > 0.1$?

If the confidence interval for δ is completely within a -0.1 to 0.1 range, then accept H_0: $|\delta| \leq 0.1$. If the confidence interval for δ is completely outside the -0.1 to 0.1 range, then accept H_1: $|\delta| > 0.1$.

5. How can a confidence interval for $\mu_1 - \mu_2$ be used to select H_0: $|\mu_1 - \mu_2| \leq 2.5$ or H_1: $|\mu_1 - \mu_2| > 2.5$?

If the confidence interval for $\mu_1 - \mu_2$ is completely within a -2.5 to 2.5 range, then accept H_0: $|\mu_1 - \mu_2| \leq 2.5$. If the confidence interval for $\mu_1 - \mu_2$ is completely outside the -2.5 to 2.5 range, then accept H_1: $|\mu_1 - \mu_2| > 2.5$.

6. How can a confidence interval for $\mu_1 - \mu_2$ be used to test H_0: $\mu_1 - \mu_2 = -5$ and then select H_1: $\mu_1 - \mu_2 > -5$ or H_2: $\mu_1 - \mu_2 < -5$?

If the lower limit for $\mu_1 - \mu_2$ is greater than -5, then reject the null hypothesis and accept H_1: $\mu_1 - \mu_2 > -5$. If the upper limit for $\mu_1 - \mu_2$ is less than -5, then reject the null hypothesis and accept H_2: $\mu_1 - \mu_2 < -5$.

7. What information can be obtained from a confidence interval for $\mu_1 - \mu_2$ that cannot be obtained from an independent-samples *t*-test?

The confidence interval for $\mu_1 - \mu_2$ provides information about the magnitude and direction of the difference while the *t*-test only provides information about the direction of the difference.

8. When would a researcher prefer to report a confidence interval for a standardized mean difference rather than an unstandardized mean difference?

If the scale of the response variable is not familiar to the intended audience, a standardized mean difference would be preferred.

9. What does it mean when a researcher declares the results of an independent-samples *t*-test to be "significant"? How should a "nonsignificant" result be interpreted?

A "significant" result simply means that the null hypothesis of identical population means has been rejected. A "nonsignificant" result should be interpreted as an inconclusive result and not as evidence that the null hypothesis is true.

10. What are the assumptions for a confidence interval or test for a difference in population means using the equal-variance method?

Random sample, independence among participants, equal variances in the two conditions, approximate normality of the response variable within each condition.

11. What are the assumptions for a confidence interval or test for a difference in population means using the unequal-variance method?

Random sample, independence among participants, approximate normality of the response variable within each condition.

12. Explain how μ_1 and μ_2 are interpreted differently in an experimental design compared to a nonexperimental design.

In an experimental design μ_1 and μ_2 describe the same study population under two different experimental situations. In a nonexperimental design μ_1 and μ_2 describe two different study populations.

13. What are the consequences of participants randomly dropping out from both groups in a two-group experiment?

A loss of power and confidence interval precision but no effect on internal or external validity.

14. What are the consequences of certain types of participants dropping out only from one group in a two-group experiment?

The internal validity of the study has been compromised and the remaining sample is no longer a random sample from the study population.

15. What are the consequences of certain types of participants dropping out equally from both groups in a two-group experiment?

The external validity of the study has been reduced and there will be a loss of power and confidence interval precision.

16. What are the advantages and disadvantages of sampling from a small study population?

It is usually easier and less costly to sample from a small study population but then the hypothesis testing and confidence interval results apply to a small study population that could have less scientific importance than a larger study population.

17. Why is a violation of the normality assumption not a major concern when testing or estimating $\mu_1 - \mu_2$ when the sample sizes are not small?

The central limit theorem tells us that the sampling distribution of the difference in sample means will be approximately normal with moderate size samples even when the response variable is highly nonnormal.

18. Why are confidence interval lines recommended in bar charts?

So that the reader will see the plausible range of population means and not just the sample means.

19. How does the value of $\tilde{\mu}_1 - \tilde{\mu}_2$ affect the sample size requirement for testing H_0: $\mu_1 = \mu_2$ with desired power?

A smaller value of $\tilde{\mu}_1 - \tilde{\mu}_2$ (the expected effects size) requires a larger sample size.

20. When would a Mann-Whitney test be preferred to an independent-samples *t*-test?

When the response variable is highly skewed and the sample size is small.

21. When would a confidence interval for a difference in population medians be preferred to a confidence interval for a difference in population means?

When the response variable is highly skewed.

22. What useful informative could be obtained by examining a confidence interval for a ratio of population MADs in a two-group design?

A ratio of population mean absolute deviations provides important information about the relative variability of the response variable in each of the two conditions.

23. A random sample of 30 students was obtained from a research participant pool of about 5,000 undergraduates at UC Santa Barbara. The sample was randomized into two groups of equal size. Group 1 received a daily meditation app and group 2 received a daily planner app. After two weeks, the Perceived Stress Scale (scored from 0 to 40 with higher scores representing greater stress) was given to all 30 participants. A 95% confidence interval for $\mu_1 - \mu_2$ was [-3.4, -0.7]. Interpret this result.

We can be 95% confident that if all 5,000 UCSB students used the daily meditation app for two weeks, their mean perceived stress score would be 0.7 to 3.4 lower than if they had all used the daily planner app for two weeks.

24. Explain why each of the following interpretations of [-3.4, -0.7] in the above example is incorrect or inadequate.

***a*) We are 95% confident that the mean stress score of the 30 students would be between 0.7 and 3.4 greater if they had used the daily planner app rather than the daily meditation app.**

The confidence interval is a statement about the population of 5,000 students and not the sample of 30 students.

***b*) We are 95% confident that the mediation app will reduce a student's stress score between 0.7 and 3.4 points.**

The confidence interval describes how the population mean of the population of 5,000 students would differ under the two treatment conditions and not how one student's score would differ.

***c*) We are 95% confident that if all 5,000 students had used the daily meditation app for two weeks, the difference in the mean stress scores would be between -3.4 and -0.7 compared to a placebo.**

Saying "the difference" does not indicate which condition is best.

***d*) We are 95% confident that $\mu_1 - \mu_2$ is between -3.4 and -0.7.**

Inadequate because μ_1 and μ_2 are not defined.

Chapter 3

1. What are the advantages of computing confidence intervals for all pairwise differences among population means instead of reporting the *p*-value for a one-way ANOVA?

The confidence intervals provide information about the order of the population mean and the differences among the population mean. The *p*-value just indicates if the null hypothesis of identical population means can be rejected.

2. What are the assumptions for the *F* test of equal population means in the one-way ANOVA?

Random sample, independence among participants, equal variances across conditions, approximate normality of response variable within each condition

3. What are the assumptions of an equal-variance confidence interval for a linear contrast of population means?

Random sample, independence among participants, equal variances across conditions, approximate normality of response variable within each condition

4. What are the assumptions of an unequal-variance confidence interval for a linear contrast of population means?

Random sample, independence among participants, approximate normality of response variable within each condition

5. What are the assumptions for a confidence interval for eta-squared?

Random sample, independence among participants, equal variances across conditions, approximate normality of response variable within each condition

6. If two-group design is analyzed using a one-way ANOVA, how will the *F*-ratio and its *p*-value compare with the equal-variance *t*-statistic and its *p*-value?

The *p*-values will be identical and the *F* value will equal the squared *t* value

7. Explain how the error variance can be reduced by including a classification factor.

If there are mean differences among any demographic groups (e.g., gender, ethnicity, etc.) within the study, those mean differences will inflate the within-group error variance. If a demographic factor is included as a classification factor, then the error variance is then defined within each treatment and demographic groups and this error variance could be substantially smaller.

8. What statistical method can be used to test H_0: $\mu_1 = \mu_2 = \mu_3$ in a between-subjects design?

one-way ANOVA

9. If the AB interaction effect in a two-factor design is large, what effects would you want to estimate and what effects may not be interesting?

Estimate the simple main effects of A at each level of B and/or the simple main effects of B at each level of A. The main effects of A and B could be misleading.

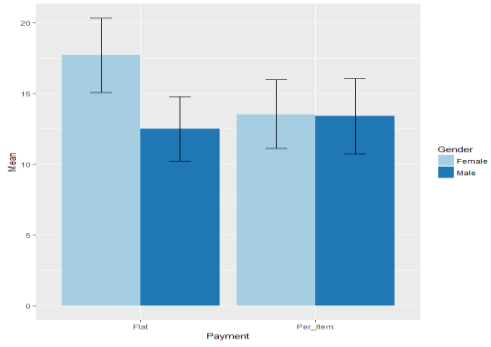
10. If the test for the AB interaction effect in a two-factor design is inconclusive, what effects would you want to estimate and what effects may not be interesting?

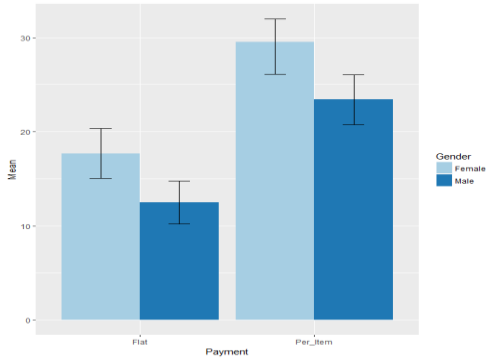
Estimate the two main effects, the simple main effects may not be interesting

11. How can you show in a convincing way that four population means are approximately equal? Explain why a "nonsignificant" one-way ANOVA result does not provide evidence of similarity.

Compute simultaneous confidence intervals for all pairwise comparisons. If all confidence intervals include 0 and are narrow, this suggests that the four population means are similar. A "nonsignificant" one-way ANOVA results could be due to low power.

12. Draw a clustered bar chart of population means in a 2 × 2 design where the interaction effect is large. Draw a second bar chart where the interaction effect is zero but one of the main effects is large.





13. Define the main effect of A in terms of the population means.

$(\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4)/2$ (.5, .5, -.5, -.5)

14. Define the main effect of B in terms of the population means.

$(\mu_1 + \mu_3)/2 - (\mu_2 + \mu_4)/2$ (.5, -.5, .5, -.5)

15. Define the AB interaction effect in terms of the population means.

$(\mu_1 - \mu_3) - (\mu_2 - \mu_4)$ or $(\mu_1 - \mu_2) - (\mu_3 - \mu_4)$ or $\mu_1 - \mu_2 - \mu_3 + \mu_4$

(1, -1, -1, 1)

16. Define the simple main effect of A at b_1 in terms of the population means.

$\mu_1 - \mu_3$ (1, 0, -1, 0)

17. Define the simple main effect of B at a_2 in terms of the population means.

$\mu_3 - \mu_4$ (0, 0, 1, -1)

18. Consider a one-way ANOVA summary table for a 4-group experiment with 10 participants per group. Fill in the missing values.

Source	*SS*	*df*	*MS*	*F*
A	210	3	70	7
ERROR	360	36	10	
TOTAL	570			

19. Suppose a researcher uses a one-way ANOVA to test H_0: $\mu_1 = \mu_2 = \mu_3$ and reports that the results are "significant". What does this mean?

It only indicates that the three population means are not all identical. It does not tell us how the means are ordered or how different they are.

20. For a 2 × 2 factorial design, give examples of the four population means that would produce the following conditions:

a) large main effect of A, zero main effect of B, zero AB interaction
b) zero main effect of A, zero main effect of B, zero AB interaction
c) large main effect of A, large main effect of B, zero AB interaction
d) zero main effect of A, zero main effect of B, large AB interaction
e) large main effect of A, zero main effect of B, large AB interaction

a)

		A	
		a_1	a_2
B	b_1	10	5
	b_2	10	5

b)

		A	
		a_1	a_2
B	b_1	10	10
	b_2	10	10

c)

		A	
		a_1	a_2
B	b_1	20	5
	b_2	20	15

d)

		A	
		a_1	a_2
B	b_1	10	5
	b_2	5	10

e)

		A	
		a_1	a_2
B	b_1	10	25
	b_2	20	15

21. Consider a two-way ANOVA summary table for a 4 × 5 factorial experiment with 5 participants per group. Fill in the missing values.

Source	*SS*	*df*	*MS*	*F*
A	15	3	5	4
B	20	4	5	4
AB	24	12	2	1.6
ERROR	100	80	1.25	
TOTAL	159			

22. If the test for an ABC interaction effect is inconclusive, what effects would you want to examine next?

The three two-way interaction effects

23. If an ABC interaction effect is detected, what effects would you want to examine next?

Simple two-way interactions or simple-simple main effects

24. If the tests for the ABC, AB, AC, and BC interaction effects are all inconclusive, what effects would you want to examine?

The three main effects

25. If an AB interaction effect is detected but the tests for the ABC, AC and BC interactions are inconclusive, what effects would you want to examine next?

The simple main effects of A at each level of B and/or the simple main effects of B at each level of A. Also, the main effect of C should be examined.

26. When would a test or confidence interval for a linear contrast of population medians be preferred to a test or linear contrast of population means?

If the response variable is highly skewed.

27. What is the distribution-free alternative to the one-way ANOVA *F* test?

Kruskal-Wallis test

28. What is an unequal-variance alternative to the one-way ANOVA *F* test?

Welch *F* test

29. Explain how you could show that an A $\times$ B interaction effect in a 2 $\times$ 2 design is non-removable.

If the confidence interval for the simple main effects of A at b_1 and A at b_2 indicate that these two population effect effects have opposite signs, or if the confidence interval for the simple main effects of B at a_1 and B at a_2 indicate that these two population effect effects have opposite signs.

30. What are the advantages and disadvantages of using a random factor rather than a fixed factor?

Confidence interval and hypothesis testing results for a random factor apply to all levels of the factor and not just the levels that were used in the study. A larger number of levels of the random factor is usually needed to obtain a usefully narrow confidence interval for the random factor effect size.

Chapter 4

1. How is a within-subjects design different from a between-subjects design?

A within-subjects design uses one group of participants and all participants are measured under all treatment conditions. A between-subjects design uses two or more groups of participants with each group being measured under only one treatment condition.

2. Explain in words how a confidence interval for $\mu_1 - \mu_2$ in a paired-samples design is computed.

Compute a difference score from the pair of scores for each participant. Compute a confidence interval for the population mean difference score.

3. How does the correlation between measurements in within-subjects designs affect the sample size requirement for a confidence interval of $\mu_1 - \mu_2$?

A smaller sample size is needed with a larger correlation.

4. What are the assumptions of the one-way within-subjects ANOVA hypothesis test for equal population means?

Random sample, independence among participants, equal variances across levels of within-subjects factor, equal correlations for all pairs of within-subjects factor levels, normal distribution of response variable within each level of the within-subjects factor.

5. What are the assumptions for a confidence interval for a linear contrast of population means in a within-subjects design?

Random sample, independence among participants, normal distribution of linear contrast scores

6. Why is a confidence interval for a linear contrast of means preferred to the test of equal population means in within-subjects designs?

Fewer assumptions (see above) and the confidence interval provide useful information about the direction and magnitude of the effect.

7. How could a confidence interval for the AB interaction effect in a 2×2 mixed design be computed?

Compute a difference score from the within-subject pair of scores for each participant and then compute a two-group group confidence interval for the difference in mean difference scores.

8. How can a confidence interval for the main effect of the between-subjects factor in a 2×2 mixed design be computed?

Compute an average of the within-subject pair of scores for each participant and then compute a two-group group confidence interval for the difference in mean average scores.

9. For a 2 × 2 within-subjects experiment, specify the linear contrast scores to estimate or test the following effects in terms of the four scores per participant (y_{11}, y_{12}, y_{21}, and y_{22}). The first subscript specifies the levels of Factor A and the second subscript specifies the levels of Factor B.

a)	A × B interaction	$y_{11} - y_{12} - y_{21} + y_{22}$
b)	main effect of A	$(y_{11} + y_{12})/2 - (y_{21} + y_{22})/2$
c)	main effect of B	$(y_{11} + y_{21})/2 - (y_{12} + y_{22})/2$
d)	simple main effect of A at b_1	$y_{11} - y_{21}$
e)	simple main effect of A at b_2	$y_{12} - y_{22}$
f)	simple main effect of B at a_1	$y_{11} - y_{12}$
g)	simple main effect of B at a_2	$y_{21} - y_{22}$

10. What is one way to control for carryover effects?

Use counterbalancing

11. What hypothesis tests can be performed using a confidence interval for $\mu_1 - \mu_2$ in a within-subjects experiment?

Directional two-sided tests, equivalence tests, noninferiority tests, and superiority tests

12. Why is a pretest-posttest design more ethical than a two-group experiment with a control group when the treatment is expected to be beneficial?

Because all participants will receive the beneficial treatment in a pretest-posttest design

13. What are some important uses of a mixed design?

The between-subjects factor will not be susceptible to carryover effects and a between-subjects classification factor can increase the generalizability of the within-subjects treatment results.

14. How does missing data affect the one-way within-subjects ANOVA vs. pairwise comparisons?

If a participant has a missing score on any of the within-subject conditions, then that participant is dropped from the one-way within-subjects ANOVA analysis. With pairwise comparisons, a participant is dropped only if this is missing data for the particular pair of conditions being analyzed.

15. What distribution-free methods could be used to analyze data from a one-way within-subjects design? When would these methods be preferred to a one-way within-subjects ANOVA or pair-wise comparisons using paired-samples *t*-tests?

Friedman test as an alternative to a one-way within-subjects ANOVA, pairwise Wilcoxon sign rank tests as an alternative to pairwise paired-samples *t*-tests, and confidence intervals for pairwise differences in medians rather than pairwise differences in means.

16. What is the effect of measurement error on the width of confidence intervals and the power of tests in between-subjects and within-subjects designs?

Greater measurement error (lower reliability) increases the widths of confidence interval and decreases the power of tests in both between-subjects and within-subjects designs but the negative effects can be more pronounced in within-subjects designs.

17. What is the effect of measurement error on estimates of δ, φ, and η^2?

Estimate of all of these measures are attenuated when the response variable contains measurement error.

18. The blood pressure measurement devices used in most behavioral labs are not highly reliable. Suppose the reliability of a single blood pressure reading is .81. What is the reliability of the average of two blood pressure measurements?

Using the Spearman-Brown formula gives: $2(.81)/(1 + (2 - 1)*.81) \approx .9$

19. Fifty UCSC freshman were randomly selected from the entering class of about 4,500 freshman and were given a public speaking examination (scored 0 to 100) in the Fall quarter of their first year (time 1). These 50 students were given the same examination during the Fall quarter of their third year (time 2). A 95% confidence interval for $\mu_1 - \mu_2$ was [-8.9, -3.7]. Interpret this result.

We are 95% confident that the mean public speaking score for all 4,500 students in their first year is 3.7 to 8.9 lower than their mean public speaking score in their third year.

20. Explain why each of the following interpretations of [-8.9, -3.7] in the above example is incorrect or inadequate.

***a*) We are 95% confident that the mean public speaking score of the 50 students increased 3.7 to 8.9 points from their first year to their third year.**

The confidence interval describes the population of 4,500 students and not the sample of 50 students.

***b*) We are 95% confident that a typical UCSC student will have a public speaking score that will improve 3.7 to 8.9 points from their third year to the third year.**

The confidence interval describes the <u>mean</u> score of the 4,500 students and not a score for <u>one</u> student.

***c*) We are 95% confident that the mean public speaking score for all 4,500 freshman drops 3.7 to 8.9 points in their third year.**

The confidence interval describes the time 1 mean minus the time 2 mean. The endpoints are negative indicating that the time 2 mean is greater than the time 1 mean.

***d*) We are 95% confident that the difference in population means is between -8.9 and -3.7.**

This is inadequate because "the difference" does not indicate if the population mean increased or decreased over time.

Appendix D. Answers to Data Analysis Problems

Problem 1-1

***a*) Describe the study population.**

The 13,800 addresses in the Watsonville database

***b*) Describe μ in the context of this study.**

μ is the mean donation amount of the 13,800 Watsonville addresses

***c*) A volunteer group has agreed to call all 13,800 addresses and ask for a donation if there is strong evidence to suggest that $\mu > 15$. Use SPSS or R to test H_0: $\mu = 15$ with $\alpha = .05$. Report *t, df,* and *p*-value in APA style and state your conclusion.**

$t(19) = 2.13$, $p = .046$. We can reject the null hypothesis and accept the alternative hypothesis that the mean donation amount of the 13,800 Watsonville addresses is greater than $15.

***d*) Use SPSS or R to compute a 95% confidence interval for μ and interpret the result.**

95% CI [15.14, 30.36]. We can be 95% confident that the mean donation amount of the 13,800 Watsonville addresses is between $15.14 and $30.36.

***e*) Hand compute a 95% confidence interval for the total dollar amount that the volunteer group should receive if they contact all 13,800 residents.**

We can be 95% confident that the total donation amount for the 13,800 Watsonville addresses is between $208,929 and $418,970.

***f*) Use the `ci.median` function in the `statpsych` package to compute a 95% confidence interval for the population median donation amount and interpret the result.**

95% CI [10, 30]. We can be 95% confident that the median donation amount of the 13,800 Watsonville addresses is between $10.00 and $30.00.

***g*) A similar study is being planned for the city of San Luis Obispo. How many residents should be contacted to obtain a 95% confidence interval for μ that has a width of $10.00? Use the sample variance (the squared standard deviation) from the Watsonville study as the planning value for the population variance.**

A random sample of 43 San Luis Obispo addresses should be contacted.

Problem 1-2

***a*) Describe the study population.**

4,000 low-income women in Oakland, CA

***b*) Describe μ in the context of this study.**

μ is the mean hours of sleep for the 4,000 low-income women in Oakland, CA

***c*) Use SPSS or R to test H_0: $\mu = 6.8$ with $\alpha = .05$. Report *t*, *df*, and *p*-value in APA style and state your conclusion.**

$t(14) = -2.94$, $p = .011$. We can reject the null hypothesis and accept the alternative hypothesis that the mean hours of sleep for the 4,000 low-income Oakland women is less than 6.8 hours per night.

***d*) Use SPSS or R to compute a 95% confidence interval for μ and interpret the result.**

95% CI [5.42, 6.58]. We can be 95% confident that the mean hours of sleep for the 4,000 low-income Oakland women is between 5.42 and 6.58 hours per night.

***e*) Use the `ci.mad` function in the `statpsych` package to compute a 95% confidence interval for the population mean absolute deviation and interpret the result.**

95% CI [0.66, 1.32]. We can be 95% confident that the mean absolute deviation in hours of sleep for the 4,000 low-income Oakland women is between 0.66 and 1.32 hours.

***f*) A similar study is being planned for Dallas, TX. How many low-income women should be sampled to obtain a 95% confidence interval for μ that has a width of 0.5? Use the sample variance (squared standard deviation) from the Oakland study as the planning value for the population variance.**

A random sample of 71 low-income Dallas women should be obtained.

Problem 1-3

***a*) Describe the study population.**

94,800 Santa Cruz county adults

***b*) Describe μ in the context of this study.**

μ is the mean critical thinking score for 94,800 Santa Cruz county adults.

***c*) Use the `ci.mean` function in the `statpsych` package to compute a 95% confidence interval for μ and interpret the result.**

95% CI [40.8, 41.4]. We are 95% confident that the mean critical thinking score for 94,800 Santa Cruz county adult is between 40.8 and 41.4.

***d*) Use the `pi.score` function in the `statpsych` package to compute a 95% prediction interval for one adult and interpret the result.**

95% PI [36.5, 45.9]. We are 95% confident that the critical thinking score for any one randomly selected Santa Cruz county adult would be between 37 and 46.

Problem 2-1

***a*) Describe the study population.**

6,000 Fresno county students enrolled in a driver's education course

***b*) Describe the population means (μ_1 and μ_2) in the context of this study.**

μ_1 is the mean driving skill score if all 6,000 students were given the training simulator program

μ_2 is the mean driving skill score if all 6,000 students were given a traditional driver's education course

***c*) Use SPSS or R to perform a directional two-sided test with α = .05. Report *t*, *df*, and *p*-value in APA style and state your conclusion. Do not assume equal population variances.**

$t(17.74) = -2.68$, $p = .015$. We can reject the null hypothesis and conclude that the mean driver skill score of the 6,000 students would be larger if they all received the simulation program than if they all received a traditional driver's education class.

***d*) Use SPSS or R to compute a 95% confidence interval for $\mu_1 - \mu_2$ and interpret the result. Do not assume equal population variances.**

95% CI [-13.03, -1.57]. We are 95% confident that if all 6,000 students were trained using the driving simulator program, their mean driving skills score would be 1.57 to 13.03 greater than if they had all been trained by the traditional method.

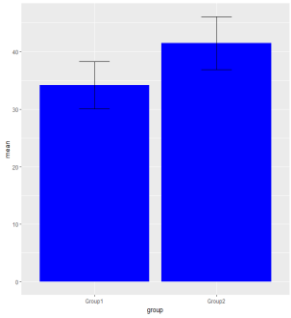
***e*) The additional simulation training will be considered superior to the traditional driver's education course if $\mu_2 - \mu_1 > 1$. Use the 95% confidence interval for $\mu_1 - \mu_2$ to decide if additional simulation training is superior to the traditional course.**

95% CI for $\mu_2 - \mu_1$ [1.57, 13.03]. The lower 95% confidence limit for $\mu_2 - \mu_1$ is 1.57 which is greater than 1 and we can conclude that the additional simulation training is superior to the traditional training method.

***f*) Use the `pi.score2` function in the `statpsych` package to compute a 95% prediction interval for the difference in scores under the two training methods for single student and interpret the result. Do not assume equal population variances.**

95% PI [-25.98, 11.39]. We can be 95% confidence that any one student in the study population of 6,000 students could score 26 points higher to 11 points lower with the driving simulator program than with the traditional training method.

***g*) Use SPSS or R to produce a bar chart for the two means with 95% confidence interval lines for the population means.**



***h*) The researcher wants to conduct a similar study in a larger city such as Chicago. How many driver education students per group (with equal sample sizes) would be needed to perform a directional two-sided test with power of .9 at α = .05 assuming a 5-point difference in population means? Use the average of the sample variances from this experiment as your planning value of the average within-group error variance.**

A random sample of 66 students should be obtained and then randomly divided into two groups of 33.

Problem 2-2

***a*) Describe the two study populations.**

The 160,000 men in the labor union directory and the 140,000 women in the labor union directory

***b*) Describe the population means (μ_1 and μ_2) in the context of this study.**

μ_1 is the mean courage score for all 160,000 men
μ_2 is the mean courage score for all 140,000 women

***c*) Use the `ci.mean2` function in the `statpsych` package to compute a 95% confidence interval for $\mu_1 - \mu_2$. Do not assume equal population variances. Interpret the result.**

95% CI [-2.73, 1.53]. We can be 95% confident that the population mean courage score for the 160,000 men is between 2.73 smaller to 1.53 larger than the population mean courage score for the 140,000 women.

***d*) Use the `ci.stdmean2` function in the `statpsych` package to compute a 95% confidence interval for the population standardized mean difference and interpret the result. Do not assume equal population variances.**

95% CI [-0.252, 0.141]. We are 95% confident that the mean courage score for the 160,000 men is 0.252 standard deviation less than to 0.141 standard deviations greater than the mean courage score for the 140,000 women.

***e*) The researchers will claim that the new courage subscale is not gender biased if they can accept H_1: $|\mu_1 - \mu_2| < 3$ with $\alpha = .05$. Use the 95% confidence interval for $\mu_1 - \mu_2$ to decide if H_1: $|\mu_1 - \mu_2| < 3$ can be accepted and state your conclusion.**

accept H_1: $|\mu_1 - \mu_2| < 3$ and conclude that the new courage subscale is not gender biased.

Problem 2-3

***a*) Describe the study population.**

The 1,000 undergraduate students in the research participation pool

***b*) Describe the population means (μ_1 and μ_2) in the context of this study.**

μ_1 is the mean willingness to ignore confession score of the 1,000 undergraduate students if they had all read vignette 1
μ_2 is the mean willingness to ignore confession score of the 1,000 undergraduate students if they had all read vignette 2

***c*) Use SPSS or R to compute a 95% confidence interval for $\mu_1 - \mu_2$. Do not assume equal population variances. Interpret the results.**

95% CI [2.01, 4.79]. We are 95% confident that the mean willingness to ignore confession score of the 1,000 undergraduate students would be 2.01 to 4.79 greater if they had all read the vignette that described the suspect as extremely thin rather than the vignette that described the suspect as tall and muscular.

***d*) Use SPSS or R to perform a directional two-sided test with α = .05. Report *t*, *df*, and *p*-value in APA style and state your conclusion. Do not assume equal population variances.**

$t(26.5) = 5.01$, $p < .001$. We can reject the null hypothesis and conclude that the mean willingness to ignore confession scores in the population of 1,000 undergraduate students would be greater if they had all read the vignette that described the suspect as extremely thin rather than the vignette that described the suspect as tall and muscular.

***e*) Use the `ci.stdmean2` function in the `statpsych` package to compute a 95% confidence interval for the standardized mean difference and interpret the result.**

95% CI [0.92, 2.72]. We are 95% confident that the mean willingness to ignore confession score of the 1,000 undergraduate students would be 0.92 to 2.72 standard deviations greater if they had all read the vignette that described the suspect as extremely thin rather than the vignette that described the suspect as tall and muscular.

f) Use SPSS or R to perform a Mann-Whitney test and report the *p*-value.

$p < .001$

***g*) The researcher wants to replicate this study using a larger sample size at a different university. How many students are needed per group (with equal sample sizes) to obtain a 95% confidence interval for $\mu_1 - \mu_2$ that has a width of 2? Use the average of the sample variances from this experiment as your planning value for the average within-group error variance.**

A random sample of 56 college students should be obtained and then randomly divided into two groups of 28 students.

Problem 3-1

***a*) Describe the study population.**

4,500 students in a research participant pool

***b*) Describe the population means (μ_1, μ_2, and μ_3) in the context of this study.**

μ_1 is the mean MPH estimates of the 4,500 students if they had all been told that the moving car bumped into the stopped car
μ_2 is the mean MPH estimates of the 4,500 students if they had all been told that the moving car crashed into the stopped car
μ_3 is the mean MPH estimates of the 4,500 students if they had all been told that the moving car smashed into the stopped car

***c*) Use SPSS or R to test H_0: $\mu_1 = \mu_2 = \mu_3$ with α = .05. Report *F*, degrees of freedom, and *p*-value in APA style.**

$F(2, 18) = 24.17$, $p < .001$

***d*) Compute unequal variance Tukey-Kramer (using R) or Games-Howell (using SPSS) pairwise confidence intervals and interpret the results.**

95% CI for $\mu_1 - \mu_2$ [-9.90, 0.47]
95% CI for $\mu_1 - \mu_3$ [-18.42, -8.43]
95% CI for $\mu_2 - \mu_3$ [-14.26, -3.17]

The results are inconclusive for the bumped vs. crashed comparison. We can be 95% confident that the mean MPH estimate of the 4,500 students would be 8.43 to 18.42 greater if they were all told "smashed" rather than "bumped" and 3.17 to 14.26 greater if they were all told "smashed" rather than "crashed".

***e*) Use SPSS or R to compute a 95% confidence interval for $\mu_3 - (\mu_1 + \mu_2)/2$ and interpret this result. Do not assume equal population variances.**

95% CI [7.29, 14.84]. We can be 95% confident that the mean MPH estimate for the 4,500 students if they were all given the smashed instruction would be 7.29 to 14.84 greater than average of the population means under the bumped and crashed instructions.

***f*) Use the `ci.lc.stdmean.bs` function in the `statpsych` package to compute a 95% confidence interval for the standardized contrast $\mu_3 - (\mu_1 + \mu_2)/2$ and interpret this result. Do not assume equal population variances.**

95% CI [1.61, 4.43]. We can be 95% confident that the mean MPH estimate for the 4,500 students if they were all given the smashed instruction would be 1.61 to 4.43 standard deviations greater than average of the population means under the bumped and crashed instructions.

***g*) Use the `ci.etasqr` function in the `statpsych` package to compute a 95% confidence interval for the population eta-squared (η^2) and interpret the result.**

95% CI [.409, .822]. We can be 95% confident that the proportion of the MPH variance in the population of 4,500 students that is predictable from the instructions is between .409 and .822.

***h*) Use SPSS to test H_0: $\mu_1 = \mu_2 = \mu_3$ with α = .05 using the Welch *F* test. Report *F*, degrees of freedom, and *p*-value in APA style.**

$F(2, 11.9) = 24.63$, $p < .001$

***i*) How many participants would be needed in a future replication of this study to obtain Bonferroni 95% confidence intervals for all pairwise differences in population means with widths of 5.0? Use the MS_E from this study as the planning value of the average within-group error variance.**

A random sample of 78 students should be obtained and then randomly divided into three groups of 26.

Problem 3-2

***a*) Describe the two study populations.**

The 2,100 male student and 2,200 female students in the university subject pool

***b*) Describe the population means (μ_{11}, μ_{12}, μ_{21}, and μ_{22}) in the context of this study.**

μ_{11} is the mean score of the 2,100 male students if they had all been tested under a flat rate condition

μ_{12} is the mean score of the 2,100 male students if they had all been tested under a per item condition

μ_{21} is the mean score of the 2,200 female students if they had all been tested under a flat rate condition

μ_{22} is the mean score of the 2,200 female students if they had all been tested under a per item condition

***c*) Test the interaction effect in a two-way ANOVA. Report *F*, degrees of freedom, and *p*-value for the interaction effect in APA style. What effects should be examined next?**

$F(1, 36) = 5.26$, $p = .028$

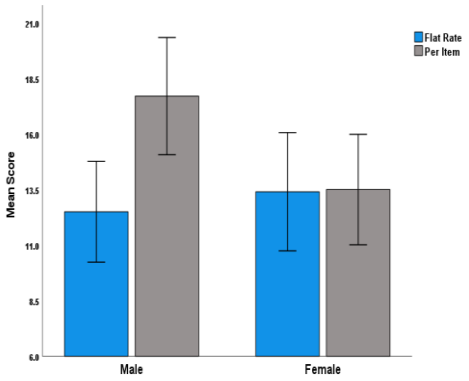
Examine the simple main effects next. For example, the simple main effect of payment type at male and female.

***d*) Compute Bonferroni 95% confidence intervals for the simple main effects of Payment at Male and at Female. Interpret the results. Do not assume equal population variances.**

95% CI at Male [-8.96, -1.43] 95% CI at Female [-4.03, 3.83]

We can be 95% confident that if all 2,100 male students were tested under the per item payment condition, their mean test score would be 1.43 to 8.96 larger than if they had all been tested under the flat rate payment condition. The effect of payment for the 2,200 female students is inconclusive.

***e*) Produce a clustered bar chart with Payment as the within-cluster factor. Include 95% confidence interval bars.**



***f*) How many participants would be needed in a future study to conduct Bonferroni tests of the two simple main effects of Payment with power of .95, $\alpha = .05$, and an effect size of 3.0? Use the MS_E from this study as a planning value of the average within-group error variance.**

A random sample of 86 male students and 86 female students should be obtained. The male students would be randomly divided into two groups of 43 and the female students would be randomly divided into two groups of 43.

Problem 3-3

***a*) Describe the study population.**

The 5,000 students in the university research participant pool

***b*) Test all main effects and interaction effects in a three-way ANOVA. For which effects can the null hypothesis be rejected and what effects have inconclusive results?**

The tests for the three-way interaction and all three two-way interaction effects are inconclusive. The null hypothesis can be rejected for all three main effects.

***c*) Compute Bonferroni 95% confidence intervals for the three main effects and interpret the results. Do not assume equal variances.**

95% CI for Style: [-46.49, -2.11] 95% CI at Size: [2.61, 46.99] 95% CI at Font: [-45.49, -1.11]

We are 95% confident that if all 5,000 students read the story in an Arial font, their mean time to read the story would be 2.11 to 46.49 seconds faster than if they had all read the story in a Times font, regardless of font size or justification.

We are 95% confident that if all 5,000 students read the story in a 12-point font, their mean time to read the story would be 2.61 to 46.99 seconds faster than if they had all read the story in a 10 point font, regardless of font type or justification.

We are 95% confident that if all 5,000 students read the story in a left justified format, their mean time to read the story would be 1.11 to 45.49 seconds faster than if they had all read the story in a full justified format, regardless of font type or font size.

***d*) The researcher wants to replicate this study using a larger sample size. About how many participants are needed in each group to obtain Bonferroni 95% confidence intervals ($\alpha^* = .05/3$) with widths of 20 seconds for the three main effects? Use the MS_E from this study as a planning value of the average within-group error variance.**

A random sample 184 students should be obtained. The 184 students would then be randomly divided into the eight treatment conditions with 23 students per condition.

Problem 4-1

***a*) Describe the study population.**

The 600 3rd year UCSC psychology majors

***b*) Describe the population means (μ_1, μ_2, and μ_3) in the context of this study.**

μ_1 is the mean satisfaction score of the 600 3rd year psychology majors for non-psychology courses
μ_2 is the mean satisfaction score of the 600 3rd year psychology majors for required psychology courses
μ_3 is the mean satisfaction score of the 600 3rd year psychology majors for elective psychology courses

***c*) Use SPSS to compute the multivariate test of H_0: $\mu_1 = \mu_2 = \mu_3$. Report *F*, degrees of freedom, and *p*-value in APA style.**

$F(2,6) = 57.64$, $p < .001$

***d*) Use SPSS or R to compute paired-samples *t*-tests for all three pairwise comparisons. Report *t*, degrees of freedom, and *p*-value in APA style for each test.**

non-psych vs. req-psych: $t(7) = -1.26$, $p = .247$
non-psych vs. elect-psych: $t(7) = -5.99$, $p = .001$
req-psych vs. elect-psych: $t(7) = -11.06$, $p < .001$

***e*) Use SPSS or R to compute Bonferroni 95% confidence interval for all three pairs of mean differences and interpret the results.**

non-psych vs. req-psych: 95% CI [-3.04, 1.29]
non-psych vs. elect-psych: 95% CI [-8.75, -2.75]
req-psych vs. elect-psych: 95% CI [-6.25, -3.50]

We are 95% confident that the mean course satisfaction rating for non-psychology courses is between 3.04 points smaller and 1.29 points larger than the mean course satisfaction rating for required psychology courses in the study population of 600 3rd year psychology students.

We are 95% confident that the mean course satisfaction rating for elective psychology courses is between 2.75 and 8.75 points larger than the mean course satisfaction rating for non- psychology courses in the study population of 600 3rd year psychology students.

We are 95% confident that the mean course satisfaction rating for elective psychology courses is between 3.50 and 6.25 points larger than the mean course satisfaction rating for required psychology courses in the study population of 600 3rd year psychology students.

***f*) Use the `ci.stdmean.ps` function in the `statpsych` package to compute Bonferroni 95% confidence interval for all three pairs of standardized mean differences and interpret the results.**

non-psych vs. req-psych: 95% CI [-0.43, 0.12]
non-psych vs. elect-psych: 95% CI [-1.52, -0.36]
req-psych vs. elect-psych: 95% CI [-1.22, -0.35]

We are 95% confident that the mean course satisfaction rating for non-psychology courses is between 0.43 standard deviations smaller and 0.12 standard deviations larger than the mean course satisfaction rating for required psychology courses in the study population of 600 3rd year psychology students.

We are 95% confident that the mean course satisfaction rating for elective psychology courses is between 0.36 and 1.52 standard deviations larger than the mean course satisfaction rating for non- psychology courses in the study population of 600 3rd year psychology students.

We are 95% confident that the mean course satisfaction rating for elective psychology courses is between 0.35 and 1.22 standard deviations larger than the mean course satisfaction rating for required psychology courses in the study population of 600 3rd year psychology students.

***g*) If the study is to be replicated at another university, how many 3rd year psychology students would be needed to obtain a 95% confidence interval for a difference in two population means that has a width of about 1.0? Use the largest variance and the smallest correlation from this study as planning values.**

Using a variance planning value of 6.667 and a correlation planning value of .919, the required number of 3rd year psychology students to sample is 19.

Problem 4-2

***a*) Describe the study population.**

A volunteer pool of 2,500 undergraduate students

***b*) Describe the four population means (μ_1, μ_2, μ_3, and μ_4) in the context of this study.**

μ_1 is the mean driving performance score of the 2,500 undergraduate students if they were tested under a light traffic and receive condition
μ_2 is the mean driving performance score of the 2,500 undergraduate students if they were tested under a light traffic and send condition
μ_3 is the mean driving performance score of the 2,500 undergraduate students if they were tested under a heavy traffic and receive condition
μ_2 is the mean driving performance score of the 2,500 undergraduate students if they were tested under a heavy traffic and send condition

***c*) Use SPSS or R to compute a test of the two-way within-subject interaction. Report *t*, degrees of freedom, and *p*-value in APA style. What effects should be examined next?**

$t(11) = 0.27$, $p = .789$. The two main effects should be examined next.

***d*) Use SPSS or R to compute Bonferroni 95% confidence intervals for the two main effects. Interpret the results.**

95% CI for Traffic: [0.34, 2.24] 95% CI for Mode: [0.01, 2.07]

We are 95% confident that the mean driving performance score for the 2,560 students would be 0.34 to 2.24 greater if they had all been tested under a light traffic condition rather than a heavy traffic condition when either sending or receiving a call. We are 95% confident that the mean driving performance score for the 2,500 students would be 0.01 to 2.07 greater if they had all been tested under a call sending condition rather than a call receiving condition in either light or heavy traffic.

Problem 4-3

***a*) Describe the study population.**

The 3,800 students in the university research participant pool

***b*) Describe the six population means ($\mu_1, \ldots, \mu_6$) in the context of this study.**

μ_1 is the mean number correct of the 3,800 undergraduate students if they were given the 3-letter word condition
μ_2 is the mean number correct of the 3,800 undergraduate students if they were given the 5-letter word condition
μ_3 is the mean number correct of the 3,800 undergraduate students if they were given the 3-letter non-word condition
μ_4 is the mean number correct of the 3,800 undergraduate students if they were given the 5-letter non-word condition
μ_5 is the mean number correct of the 3,800 undergraduate students if they were given the 3-digit number condition
μ_6 is the mean number correct of the 3,800 undergraduate students if they were given the 5-digit number condition

***c*) Use SPSS or R to test the string length by string type interaction (this can be done using a two-way mixed ANOVA or a one-way ANOVA on the difference scores). Report *F*, degrees of freedom, and *p*-value in APA style for the interaction effect. What effects should be examined next?**

$F(2, 15) = 11.92$, $p < .001$. The simple main effects of string length at words, non-words, and numbers could be examined next.

***d*) Use SPSS or R to compute Bonferroni 95% confidence intervals for the simple main effects of string length at words, non-words, and numbers. Interpret the results.**

95% CI for length at word: [-0.49, 3.49]
95% CI for length at non-word: [2.18, 5.82]
95% CI for length at number: [0.09, 1.91]

The results are inconclusive for word. We are 95% confident that the mean number of correct answers for the 3,800 students would be 2.18 to 5.82 greater with a 3-letter non-word than a 5-letter non-word. We are 95% confident that the mean number of correct answers for the 3,800 students would be 0.09 to 1.91 greater with a 3-digit number than a 5-digit number.

Appendix E. Probability Rules

The uncertainty of some "event" can be quantified on a 0 to 1 *probability scale* where a probability of 0 indicates that the event cannot occur and a probability of 1 indicates that the event must occur. An event can be any clearly defined outcome such as "a flipped coin will land heads" or "it will rain tomorrow". In statistical applications, we can define events in random samples such as "a confidence interval will include the value of the population parameter" or "a hypothesis test will result in an incorrect decision". In any one random sample that we obtain, we can compute a confidence interval or perform a hypothesis test. Although we cannot know with complete certainty if a computed confidence interval includes the unknown population parameter value or if a hypothesis test has resulted in an incorrect conclusion, we can use probability to quantify our degree of uncertainty regarding confidence interval or hypothesis testing results from a single sample. The subjective definition of probability, rather than the relative frequency definition (see section 1.17), must be used to quantify our uncertainty regarding confidence interval or hypothesis testing results that were observed in any one sample. The probability rules in this appendix are general and apply to both the relative frequency and subjective definitions of probability.

Notation and Definitions

- E resents some event
- $P(\mathrm{E})$ is the probability that event E will occur
- $P(\mathrm{E}^{\mathrm{C}})$ is the probability that event E will not occur
- $P(\mathrm{E}_1$ *or* $\mathrm{E}_2)$ is the probability that event E_1 will occur or event E_2 will occur
- $P(\mathrm{E}_1$ *and* $\mathrm{E}_2)$ is the probability that events E_1 and E_2 both will occur
- $P(\mathrm{E}_1|\mathrm{E}_2)$ is the *conditional probability* that event E_1 will occur given that event E_2 already has occurred
- If $P(\mathrm{E}_1$ *and* $\mathrm{E}_2) = 0$ then E_1 and E_2 are *mutually exclusive events*
- If $P(\mathrm{E}_1$ *and* $\mathrm{E}_2) = P(\mathrm{E}_1) \times P(\mathrm{E}_2)$ then E_1 and E_2 are *independent events;* otherwise, the two events are *dependent*

Some Basic Probability Rules

Rule 1. $0 \leq P(\mathrm{E}) \leq 1$

This rule states that a probability can be any value from 0 to 1 where $P(\mathrm{E}) = 0$ indicates that the event cannot occur and $P(\mathrm{E}) = 1$ indicates that the event must occur.

Rule 2 (Complement rule). $P(\mathrm{E}^{\mathrm{C}}) = 1 - P(\mathrm{E})$

Example. If the probability that a confidence interval for μ will include the value of μ is .95, then the probability that the confidence interval will not include the value of μ is 1 – .95 = .05.

Rule 3. If $P(\mathrm{E}) \leq p$ then $P(\mathrm{E}^{\mathrm{C}}) > p$

Example. If the probability that a confidence interval will include the value of a population parameter is greater than or equal to .95, then the probability that it will not include the value of the population parameter is less than .05.

Rule 4 (Addition rule). $P(\mathrm{E}_1 \textit{ or } \mathrm{E}_2) = P(\mathrm{E}_1) + P(\mathrm{E}_2) - P(\mathrm{E}_1 \textit{ and } \mathrm{E}_2)$

Note: If $P(\mathrm{E}_1 \textit{ and } \mathrm{E}_2)$ is unknown but E_1 and E_2 are assumed to be independent events, then $P(\mathrm{E}_1 \textit{ and } \mathrm{E}_2)$ can be set equal to $P(\mathrm{E}_1) \times P(\mathrm{E}_2)$.

Example. Let $P(\mathrm{E}_1) = .025$ represent the probability that a directional two-sided test in an original study resulted in a directional error, and let $P(\mathrm{E}_2) = .025$ represent the probability that a directional two-sided test in a replication of the original study resulted in a directional error. The two events are assumed to be independent. The probability that a directional error occurred in the original study or the replication study is $.025 + .025 - .025^2 = .049375$.

Rule 5 (Multiplication rule). $P(\mathrm{E}_1 \textit{ and } \mathrm{E}_2) = P(\mathrm{E}_1) \times P(\mathrm{E}_2|\mathrm{E}_1) = P(\mathrm{E}_2) \times P(\mathrm{E}_1|\mathrm{E}_2)$

Note: If E_1 and E_2 are independent events, then $P(\mathrm{E}_2|\mathrm{E}_1) = P(\mathrm{E}_2)$, $P(\mathrm{E}_1|\mathrm{E}_2) = P(\mathrm{E}_1)$, and $P(\mathrm{E}_1 \textit{ and } \mathrm{E}_2) = P(\mathrm{E}_1) \times P(\mathrm{E}_2)$.

Note: If $\mathrm{E}_1, \mathrm{E}_2, \ldots \mathrm{E}_k$ are all independent events, then

$$P(\mathrm{E}_1 \textit{ and } \mathrm{E}_2 \textit{ and } \ldots \textit{ and } \mathrm{E}_k) = P(\mathrm{E}_1) \times P(\mathrm{E}_2) \times \ldots \times P(\mathrm{E}_k).$$

Example. Let $P(\mathrm{E}_1) = .95$ represent the probability that a 95% confidence interval for a mean difference in a sample of male participants includes the value of the male population mean difference, and let $P(\mathrm{E}_1) = .95$ represent the probability that a 95% confidence interval for a mean difference in a sample of female participants includes the value of the female population mean difference. The two events are assumed to be independent. The probability that both confidence intervals include the values of their population mean differences is .95 x .95 = .9025.

Rule 6. If $\mathrm{E}_1, \mathrm{E}_2, \ldots, \mathrm{E}_k$ are mutually exclusive events then

$$P(\mathrm{E}_1 \textit{ or } \mathrm{E}_2 \textit{ or } \ldots \textit{ or } \mathrm{E}_k) = P(\mathrm{E}_1) + P(\mathrm{E}_2) + \ldots + P(\mathrm{E}_k)$$

Example. A confidence interval will be used to select one of the following three hypotheses: 1) H_1: $\mu_1 - \mu_2 < 10$, 2) H_2: $\mu_1 - \mu_2 > 10$, or 3), or H_3: $10 \leq \mu_1 - \mu_2 \leq 10$. Suppose $P(\mathrm{E}_1) = .01$ is the probability that the confidence interval will support H_1: $\mu_1 - \mu_2 < 10$, $P(\mathrm{E}_2) = .01$ is the probability that the confidence interval will support H_2: $\mu_1 - \mu_2 > 10$, and $P(\mathrm{E}_3) = .98$ is the probability that the confidence interval will support H_3: $10 \leq \mu_1 - \mu_2 \leq 10$. The support of H_1 and the support of H_2 are the two events of interest. The probability that the confidence interval will support H_1 or H_2 is .01 + .01 = .02.

Rule 7 (Boole's inequality). If $\mathrm{E}_1, \mathrm{E}_2, \ldots, \mathrm{E}_k$ are not necessarily mutually exclusive events then

$$P(\mathrm{E}_1 \textit{ or } \mathrm{E}_2 \textit{ or } \ldots \textit{ or } \mathrm{E}_k) \leq P(\mathrm{E}_1) + P(\mathrm{E}_2) + \ldots + P(\mathrm{E}_k)$$

Note: Boole's inequality is a useful approximation to $P(\mathrm{E}_1 \textit{ or } \mathrm{E}_2 \textit{ or } \ldots \textit{ or } \mathrm{E}_k)$ only if each $P(\mathrm{E}_j)$ is small.

Example. Suppose two ($k = 2$) hypothesis tests are computed from the same sample. Let E_1 represent the event that the first directional two-sided test will result in a directional error and let E_2 represent the event that the second directional two-sided test will result in a directional error. Events E_1 and E_2 are not mutually exclusive because it is possible that both hypothesis tests will result in directional errors. Since $P(\mathrm{E}_1) = .05$ and $P(\mathrm{E}_2) = .05$, then $P(\mathrm{E}_1 \textit{ or } \mathrm{E}_2) \leq .10$.

Rule 8 (Bonferroni's inequality). If $\mathrm{E}_1, \mathrm{E}_2, \ldots, \mathrm{E}_k$ are events (independent or dependent) then

$$P(\mathrm{E}_1 \textit{ and } \mathrm{E}_2 \textit{ and } \ldots \textit{ and } \mathrm{E}_k) \geq P(\mathrm{E}_1) + P(\mathrm{E}_2) + \ldots + P(\mathrm{E}_k) - k + 1$$

Note: Bonferroni's inequality is a useful approximation to $P(\mathrm{E}_1$ *and* E_2 *and* … *and* $\mathrm{E}_k)$ only if each $P(\mathrm{E}_j)$ is large.

Example. Suppose two (k = 2) confidence intervals for differences in population means are computed from the same sample. Let E_1 represent the event that a 95% confidence interval for of $\mu_1 - \mu_2$ includes the value of $\mu_1 - \mu_2$ and let E_2 represent the event that a 95% confidence interval for $\mu_2 - \mu_3$ includes the value of $\mu_2 - \mu_3$. Since $P(\mathrm{E}_1) = .95$ and $P(\mathrm{E}_2) = .95$, then $P(\mathrm{E}_1$ *and* $\mathrm{E}_2) \geq .90$.

Rule 9 (rule of total probability). For $\mathrm{E}_1, \mathrm{E}_2, \ldots, \mathrm{E}_k$ mutually exclusive events, if $P(\mathrm{A}\,|\,\mathrm{E}_1), P(\mathrm{A}\,|\,\mathrm{E}_2), \ldots, P(\mathrm{A}\,|\,\mathrm{E}_k)$ are known and $P(\mathrm{E}_1), P(\mathrm{E}_2), \ldots, P(\mathrm{E}_k)$ are known, then

$$P(A) = P(\mathrm{A}\,|\,\mathrm{E}_1)\,P(\mathrm{E}_1) + P(\mathrm{A}\,|\,\mathrm{E}_2)\,P(\mathrm{E}_2) + \ldots + P(\mathrm{A}\,|\,\mathrm{E}_k)\,P(\mathrm{E}_k)$$

Example. Student job placement probabilities have been determined within each of four divisions of a college. These probabilities are .45, .68, .59, and .80. The probabilities of a student being in one of the four divisions are .35, .28, .15, and .22. Given this information, the probability that a student from this college gets a job placement is .45(.35) + .68(.28) + .59(.15) + .80(.22) = .6124.

Rule 10 (Bayes' rule). If $P(\mathrm{A}\,|\,\mathrm{B})$, $P(\mathrm{A})$, and $P(\mathrm{B})$ are known, then $P(\mathrm{B}\,|\,\mathrm{A}) = P(\mathrm{A}\,|\,\mathrm{B})P(\mathrm{B})/P(\mathrm{A})$

Example. A new test for childhood autism was given to a sample of children known to have autism and 89% tested positive (this is called the "specificity" of the test). The test was also given to a general population of children and 10% tested positive. Experts believe that about 3% of all children have autism. The probability that a randomly selected child has autism after testing positive is .89(.03)/.10 = .267.

Note 1: This version of Bayes' rule is most useful in applications where the required probabilities have been obtained from difference sources and are then pieced together to compute the required conditional probability. If $P(\mathrm{B}\,|\,\mathrm{A})$, $P(\mathrm{A})$, and $P(\mathrm{B})$ have been estimated from a single study, the data used to estimate these probabilities could be summarized in a *contingency table* (see Chapter 3 of Volume 3). The data in a contingency table can then be used to compute any conditional probability without the use of Bayes' rule.

Note 2: Another version of Bayes' rule replaces $P(\mathrm{A})$ with an equivalent value given by the rule of total probability.