2021

Indispensable Statistics for the Behavioral Sciences ~With SPSS 26

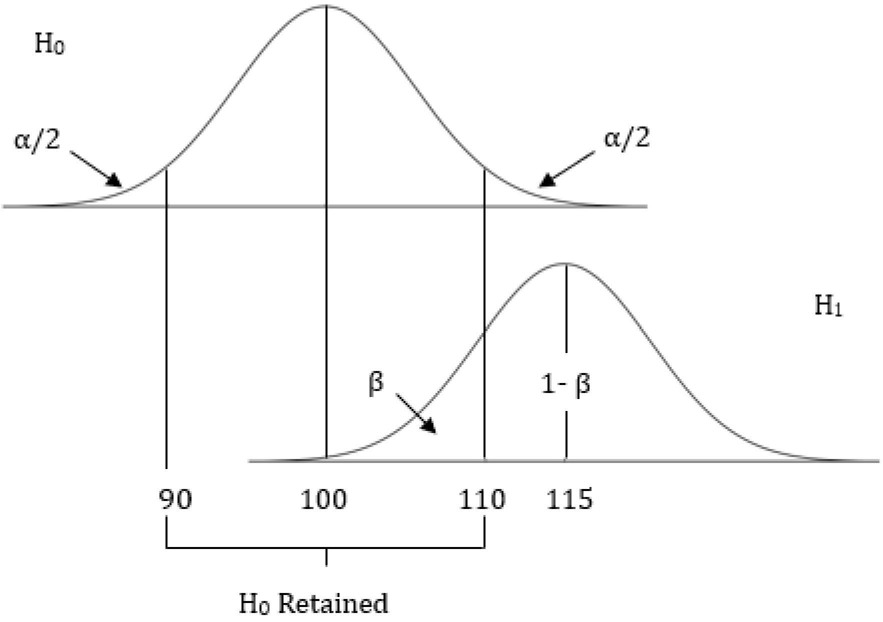
Howard Reid Ph.D.
Buffalo State College, reidhm@buffalostate.edu

Indispensable Statistics for the Behavioral Sciences ~With SPSS 26~

By

Howard M. Reid, Ph.D.

SUNY Buffalo State



Spring 2021

Dedication

This book is dedicated to the memory of my father, Malcolm, and my brother, Robert. Tragically, each died much too young. Their passion for life and their devotion to family and friends are greatly missed by all who had the good fortune to know them. And this book is dedicated to the memory of my mother, Jeannette, whose love, support, and strength I always found to be a blessing.

Brief Contents

TABLE OF CONTENTS

About the Author

Howard M. Reid is a professor at SUNY Buffalo State. He received his doctorate in experimental psychology from the University of Maine followed by postdoctoral study in behavioral genetics at The Jackson Laboratory in Bar Harbor, Maine. He has been an active researcher with broad interests, including published work in operant analysis of behavior, animal models of epilepsy, links between ADHD (attention-deficit/hyperactivity disorder) and laterality, as well as construction of a scale measuring student appreciation of the liberal arts. He has also directed numerous undergraduate research projects and has been active in shared governance, most notably chairing his college's senate seven times. He is the recipient of a number of prestigious awards including the SUNY Chancellor's Awards in Teaching as well as Faculty Service, the Buffalo State President's Award for Excellence as an Undergraduate Research Mentor and, most recently, the college's Career Achievement Award.

Preface

"Without data you're just another person with an opinion."

W. Edwards Deming

While there are many fine introductory statistics books, undergraduate students often continue to view statistics courses negatively. And many fear they will be unable to master the basic level of understanding that is essential to progress in their majors. The present text is an attempt to rethink what students majoring in the behavioral sciences absolutely must learn in an introductory statistics course and how best to organize the presentation of this material so they can succeed in their chosen field of study.

Every book is written from some perspective. The perspective of this book is that a first course in applied statistics is an introduction to a form of critical thinking as much as it is an introduction to a series of mathematically-based procedures. And this book emphasizes what a student will need to remember semesters, even years, in the future rather than focusing upon a cursory introduction to numerous techniques, many of which will soon be forgotten. Finally, this book is designed to provide a foundation upon which students can build if they take further statistics or methodology courses.

As a consequence, while this book covers many of the same topics as other texts, the presentation, and in some cases the content, differs in significant ways. First, the text is organized to assist students in understanding the logic of statistical procedures and how these procedures are related to each other. And the order the material is presented has been chosen so that students gain confidence in their ability to master this subject. For instance, the mathematically less challenging procedures that are employed with nominal data are presented before the more involved procedures that are used with interval and ratio data. And concepts build upon earlier material so that by the end of the book readers will have gained a clear comprehension of the goals, basic techniques and the limitations of statistical analysis. Second, an aim of the text is to have the reader not only be exposed to fundamental concepts such as variability, but also to come to appreciate that these concepts re-occur in a variety of contexts. Repetition and an emphasis upon the use of definitional equations enhance gaining a deeper understanding. A third major thrust is that this text emphasizes the mastery of SPSS through the use of step-by-step directions and numerous figures, all integrated with the statistical procedures being learned. As a result, students see how mastery of SPSS complements their learning of statistical procedures. Fourth, a concerted effort has been made to integrate the study of statistics with numerous disciplines as well as by including brief historical sections and incorporating relevant quotations. While it is unlikely that each of these will resonate with every reader, the goal is to present statistics in a manner that appeals to students with diverse backgrounds and interests.

The result is, I believe, a text that will assist students in appreciating the value of statistical analyses as well as mastering a set of commonly employed statistical procedures. By following a logical progression of topics, focusing upon just the statistical concepts and procedures that are absolutely indispensable for students to master, and presenting the material in an easy to read manner, this text is enhancing the success of my students. And many actually come to enjoy the material.

Acknowledgements

A list of everyone whose contribution to this book could appropriately be acknowledged would be quite long. However, the contributions of a number of individuals were so significant that I would be particularly remiss if they were not recognized:

First, I want to especially thank my wife, Dr. Susan E. Mason, whose wisdom and support I have always found to be invaluable.

I wish to also thank Dr. Robert L. Collins who first introduced me to the field of statistics, and who has taught me so much over many years.

I want to additionally recognize the contributions of Drs. Dwight Hennessy, Jill M. Norvilitis and Karen O'Quin, friends and colleagues to whom I have often turned for advice.

Three former SUNY Buffalo State undergraduates, Leanna Kalinowski, Aveary Menze, and Morgan Morningstar are sincerely thanked for their assistance with the development of the book and the PowerPoint slides that accompany this text.

I also wish to thank Meghan Pereira, Senior Instructional Designer, and Neil O'Donnell, Senior Counselor, for providing essential support in the presentation of this text.

In addition, two very supportive secretaries, Ms. Kelly Boos and Ms. Karen Skoney, deserve special recognition.

I also gratefully acknowledge the very significant contribution of Sean Clark. Following his graduation with dual majors in Psychology and Computer Information Systems, Sean provided invaluable assistance in creating the figures in this text using Python.

Finally, I want to thank my many statistics students. Their questions and feedback led to innumerable improvements in this text.

INTRODUCTION

Chapter 1 – Why You Need to Understand Statistical Reasoning

Chapter 1
Why You Need To Understand Statistical Reasoning

"Few things mislead us more than failure to grasp simple statistical principles."

Sharon Begley

Introduction

Statistics is a general term that has a number of meanings and uses. For instance, statistics can refer to a variety of procedures that have been found to be helpful, even essential, in fields such as psychology, economics, sociology, and political science, as well as many others. In addition, newspapers are filled with statistical summaries. Politicians turn to statistics for support of their policies as well as to attack their opponents. And what would sports be without statistical analyses? Our high standard of living would not be possible without the use of statistics in business and finance. Our health has been immeasurably enhanced by employing statistical procedures. And, when we die, though it is a time of sadness and reflection for family and friends, we will become another statistic in census books. You cannot avoid statistics and you should not try to. Statistics are useful and ubiquitous.

Overview and Mathematics Review

Statistics is an example of a field that has had an impact far beyond what any of its early founders could have imagined. Before the late 1800s there was no recognizable discipline of statistics. Now, knowledge of statistical procedures is seen as essential for undergraduates majoring in any of the natural or social sciences, and specialized statistics courses are offered in most college mathematics departments. For many careers a knowledge of statistics has become essential.

Rudimentary forms of statistical analysis began many centuries ago with empires that were interested in measuring trade and enhancing efficient taxation. This use of statistics continued and gradually expanded. In the 16th and 17th centuries early studies of probability provided a new perspective. For example, Gerolamo Cardana (1501–1576), in the first book on probability, noted that if a die (singular of dice) is fair then each face will have an equal chance of occurring. And Jacques Bernoulli (1654–1705) showed that the greater the number of times a die is tossed, the more closely the results come to the predicted probabilities. This understanding of probability was first put to practical use by wealthy aristocrats who wanted to optimize their likelihood of success in games of chance.

The use of statistics continued to expand, and the knowledge of the field grew. For instance, John Graunt (1620–1674) estimated the population of London based upon mortality figures in a 1662 book entitled 'Natural and Political Observations Upon Bills of Mortality'. He checked his predictions by collecting data from three parishes, which was the first instance of representative sampling. And during the late 18th century Pierre-Simon Laplace (1749-1827) described what is now known as the normal distribution. Much of this text focuses upon learning to use procedures that are appropriate for data that are 'normal'. Also, William Playfair (1759-1823) introduced the bar chart, histogram and pie chart to summarize data. We will be reviewing these presentations of data in Chapters 2 and 3.

The field of statistics expanded dramatically in the late 19th and early 20th centuries. Florence Nightingale (1820–1910), a widely known reformer, employed statistics to support her campaign to improve hospital care. Francis Galton (1822–1911) introduced the concept of the correlation, though it was his student Karl Pearson (1857–1936) who developed the equation that is used to this day. In honor of Pearson's accomplishment, this procedure is known as the Pearson correlation. You will learn about this procedure in Chapter 14. Pearson also developed what is known as the chi-square (pronounced ki square) goodness-of-fit test, the topic of Chapter 7, and was the first to use the term 'standard deviation', a concept reviewed in Chapter 3. During the same period, William Gosset (1876–1937) proposed the t test, a procedure covered in Chapters 9 and 10, to assist in making decisions based upon small samples. At the time, Gosset worked for the Guinness brewery in Dublin and was prevented from publishing the procedure under his own name. As a result, he published his paper using the name 'Student', and to this day the procedure is sometimes referred to as Student's t test. Somewhat later, Sir Ronald Fisher (1890–1962) made numerous contributions to the field of statistics, but he is probably most famous for his foundational work leading to the Analysis of Variance (ANOVA). Much of the latter part of this book (Chapters 11 – 13) is devoted to describing this set of useful procedures.

While this review of the history of statistics is obviously brief, no survey of statistics should omit mentioning Egon Pearson (1895-1980, the son of Karl Pearson) and Jerzy Neyman (1894-1981). Among their contributions are an emphasis upon Type II error and power, concepts which are introduced in Chapter 6, as well as a focus upon confidence intervals, which are first discussed in Chapter 9.

The Definition Of Statistics

We have seen in our brief introduction that statistics have been found to be useful in a wide variety of fields and that there have been numerous contributors to the growth of this discipline.

But what, exactly, is it? The quickest answer is to provide a definition. A common definition indicates that the term statistics is used in two ways. First, it is "the mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling". And second, it is used as a shorthand for "numerical data". This is a complex definition. And you may not feel that it has provided much clarification. This may be due to the inclusion of a number of terms that are probably somewhat familiar, but you may not know their precise meanings. For instance, the above definition assumes you know the meaning of words such as data, population, inference and sampling. Do not be concerned if your understanding of any (or even all) of these terms is somewhat 'hazy'. We will be reviewing the definitions of each of these terms, as well as many others, as we progress through the text. We begin by making a fundamental distinction.

The Two Uses Of Statistics

As you will see, there are numerous statistical procedures. Fortunately, however, each is used in one of two related ways. First, they assist us in seeing the world around us more clearly. And, second, they are an aid in thinking more accurately.

Seeing Clearly Can Be Harder Than It Appears

We tend to be confident that we can rely upon our senses. After all, we are able to recognize our friends, we can drive our cars without hitting anything, and we can tell when someone is happy or sad. Unfortunately, while each of these statements is usually true, we all know that there are exceptions. Sometimes we have difficulty recognizing even people we know well if we meet them in a novel situation. Sometimes, due to rain, snow or fog, we cannot see well enough to drive safely. And sometimes we are mistaken when we try to read another person's emotions.

A famous example of inaccurate perception is what psychologists call the Attractiveness Bias – we have a tendency to attribute good qualities to attractive people and bad qualities to unattractive people. This is certainly not a new situation, for the witches in folktales are always ugly, and the hero and heroine are attractive. What you may not be aware of is how pervasive this bias is. For instance, research has indicated that attractive defendants are less likely to be found guilty by a jury and, if they are convicted, they receive lighter sentences! That is wonderful news for those of you who are good looking, but not helpful for the rest of us.

Thinking Clearly is Also Harder Than It Appears

Psychologists have also found that people don't think as accurately as they suppose they do. For instance, when people estimate risks they frequently rely upon personal experience, and especially recent news reports. Thus, when asked which is more dangerous, taking a trip by flying in a plane or riding in a car, many people base their judgment on what they have heard in the news and respond that flying in a plane is more dangerous. Actually, riding in a car is much more dangerous, but car accidents are less likely to receive spectacular coverage in the news than aircraft mishaps. This is known as the Availability Bias.

Knowledge Of Statistics Can Help You See And Think More Accurately

We have mentioned only a few examples of not seeing or thinking effectively. Cognitive scientists have demonstrated many ways that we tend to make errors. Fortunately, there are a variety of statistical procedures to help us see more clearly and think more accurately. Since no one wants to be wrong, these are very valuable techniques indeed.

So why aren't you thinking statistically now? Perhaps you *never had an opportunity to learn to use statistics*. Then this is your lucky day! Your study of statistics will have many rewards. Perhaps you have avoided learning statistics because you think that *statistics are scary*. Actually the vast majority of students are able to master the statistical procedures covered in this text in one semester, and many even come to appreciate the logic of these procedures. Of course, there are lots of numbers, but if you apply yourself you will do fine. You may also fear that *statistics are boring*. However, since the procedures you will be learning will help you gain a better understanding of the world and the people in it, statistics will make your life more interesting, not less.

Plan Of This Book

This book is designed to assist you in *using statistics, first to see more clearly and then to think more accurately*. Using statistics to see information more clearly is called **descriptive statistics**. Specifically, descriptive statistics include a number of techniques that enable you to summarize a set of observations or numbers so they are easily understood. In the study of statistics, factual information, often in the form of numbers, is called **data** (plural of datum). Thus, the early part of this book will teach you procedures that will enable you to 'see' a set of data more accurately.

Descriptive statistics – Techniques that are used to summarize data. These procedures lead to a better understanding of the data.

Data (plural of datum) – Observations or factual information, often in the form of numbers.

The remainder of the book will show you how to use statistical procedures to think more clearly about data. This is called **inferential statistics**. When you make an inference, you are making a decision based upon your data. Specifically, in the part of the book covering inferential statistics you will learn a set of procedures which will assist you in making better decisions. For instance, you will discover how we can be confident with the conclusion that attractive people are treated more favorably by juries than are less attractive people.

Inferential statistics – Techniques that are used in making decisions based upon data.

Descriptive statistics and inferential statistics – that's all there is to this book. What could be simpler?

Goal Of This Book

In life most of us find that it is a good idea to have both *breadth and depth of knowledge*. It is important to know some area or skill well (depth), but it is also important to have some knowledge of many things (breadth). For instance, in college you select a major that will provide you with the depth of knowledge necessary for your chosen career. At the same time you are often expected to take a variety of other courses to broaden your education. It is the goal of this book to introduce you to both the breadth and depth of statistics. As with home or auto repair, you will need to be able to pick the right 'tool' among many that are available (breadth), and know how to use it well (depth).

Taking One Step At A Time

The techniques and logic of statistics cannot be mastered all at once. Instead, this book is designed as a progressive process so that you can learn statistics in as efficient a manner as is possible. It is also my goal to make this process reasonably enjoyable. So let's get started!!!

Progress Check

1. The two uses of statistics assist us in _____ and _____ more clearly.
2. Using statistics to see more clearly is called _____ statistics.
3. Using statistics to assist us in making decisions is called _____ statistics.

Answers: 1. seeing; thinking 2. descriptive 3. inferential

A Brief Review Of Mathematics

This book's goal is to assist you in understanding and learning how to use statistics. Thus, you will be reading about a variety of procedures that have been found to be useful in a range of settings. The book will not be explaining how these procedures were developed or why they take the specific form that they do. Thus this book, though mathematical, will likely have a very different focus from any mathematical book you have previously read. Obviously for most people, including me, the manipulation of numbers is not a favorite pastime. In fact, it is even possible that at least some readers may find mathematics to be aversive, and you may be concerned about the prospect of learning about this mathematical discipline. If you are one of these readers you should know that it is a major goal of this text to alter that view. By the end of this book you should feel a sense of pride in mastering an introduction to a mathematical discipline and, further, you should have gained an appreciation for the usefulness of this field. You may even find that your view of mathematics in general has changed!

No doubt about it, statistics is mathematical. If you leaf through the pages of this, or any, statistics text, you will see graphs, tables filled with numbers, and some very impressive-looking equations. You may even feel intimidated. Do not be. You will see that the only background in mathematics you need is knowledge through basic algebra and the ability to use graphs. Since your mathematics may be quite 'rusty' we will now turn to a brief review of the procedures required for this book. You will probably be relieved to learn that they are quite modest.

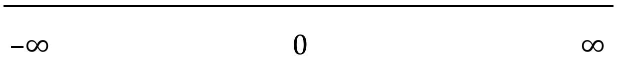
The Need For Math To Follow Rules

Numbers are simply symbols. They take the place of something else. For instance, if you were asked how many newspapers or magazines you have read in the last year, by using mathematics you do not have to actually produce each of them. Instead, you just state a number. Clearly, saying a number is much more convenient than carting around a load of old newspapers. Further, just as we can manipulate another type of symbol, the letters of the alphabet, to make words, we can also manipulate numbers. In both cases, however, we must follow rules in order to be understood. It would not be helpful to you if I had used my own, unique system for spelling when this book was written. Similarly, it would not be helpful if each of us followed our own unique system for manipulating numbers. There have to be some agreed-upon rules. Fortunately, to master the material of this book there are surprisingly few mathematical rules. However, they will be used repeatedly and thus it is critical that you understand them.

Numbers frequently signify the magnitude of something. For example, we all know that 10 apples are more than 5 apples. This concept of magnitude can be linked to what is called a number line. A number line is simply a line upon which all possible numbers can be located. It stretches from negative infinity ($-\infty$) through 0 to positive infinity ($+\infty$, commonly symbolized as just ∞) which is shown in Figure 1.1. It should be clear that the farther to the right of 0 you go, the larger the positive number and the farther to the left of 0 you go, the larger the negative number. And, for every possible positive number there is a corresponding negative number. The line is, therefore, symmetrical. When a pair of numbers that are identical, except for their sign, are added together, the result is 0. For instance, the sum of +7 and –7 is 0. And with the number line it is clear that the result of adding +10 with –4 equals +6, or just 6. And if you sum –7 and +3, the answer would be –4.

You should also remember that if you multiply or divide two positive numbers, the result is a positive number. Thus, 8 X 4, which can also be written as (8)(4), equals 32, and 6 ÷ 3 equals 2. If you multiply or divide two negative numbers, the result is also positive. For instance, (–3) X (–4) equals 12, not –12, and (–8) ÷ (–4) equals 2. Finally, the result of multiplying or dividing one number that is positive and another number that is negative is a negative number. Thus, (–3) X (2) is equal to –6, not 6, and (–4) ÷ (2) equals –2.

Figure 1.1 The Number Line



You also need to understand that when you square a number, you are multiplying the number by itself. Thus, 7 squared, which is written 7^2, is equivalent to 7 X 7 which equals 49, and it is important for you to remember that $(-4)^2$ equals 16, not –16. Similarly, the square root of a number is the number that when multiplied by itself would equal the number with which we are concerned. This course will only be dealing with the positive square roots. Therefore, the square root of 49, which is written $\sqrt{49}$, is equivalent to 7. To check this, 7 X 7 equals 49. As we will frequently be squaring and finding the square root of numbers, it is essential that you have a calculator with these functions. However, to use this book it is not necessary that you have a sophisticated statistical calculator.

Two additional mathematical concepts that you will be using are the inequalities 'less than' and 'greater than'. These concepts are symbolized by < and >, respectively. Writing $X < 10$ indicates that the value of X must be less than 10. It might be 9.99 or 0 or –20 or any other number less than 10. We do not know the precise value of X, but we know that it must be less than 10. Similarly, $X > 4$ indicates that X must have a value greater than 4. We do not know the precise value, but we do know that it cannot be 4 or less than 4.

You also must be familiar with proportions and percentages. For instance, a proportion of .50 is equal to one half, which is equivalent to 50%. And a proportion of .10 is equal to one tenth, which is equivalent to 10%. In addition, it should be obvious that if there were 100 people and the proportion who had visited Europe was .25, then the number of these people who had visited Europe is 25. This is shown with the equation .25 X 100 = 25. Similarly, if the number of people was 1000, then the number of these people who had visited Europe would be 250. This is shown with the equation .25 X 1000 = 250.

The next mathematical concept in this brief review is the **absolute value**. The absolute value is the magnitude of a number irrespective of whether it is positive or negative. Thus the absolute value of -3 is 3. And the absolute value of $+3$ is also 3.

<u>Absolute value</u> – The magnitude of a number irrespective of whether it is positive or negative.

Finally, it is essential that you remember the order in which mathematical operations are performed: begin with items within parentheses, then exponents and roots, then multiplication and division, and lastly addition and subtraction (Table 1.1). Thus $(2 + 3)^2$ is equal to 5^2 which is 25, but $2 + 3^2$ is equal to $2 + 9$ which is equal to 11. And you proceed from left to right. For example, remembering that we are only dealing with positive square roots, we find:

$\sqrt{}(5^2 – 16) – 4$

$= \sqrt{}(25 – 16) – 4$

$= \sqrt{}9 – 4$

$= 3 – 4$

$= –1.$

Table 1.1 Order of Mathematical Operations

1	Operations within parentheses
2	Exponents and roots
3	Multiplication and division
4	Addition and subtraction

And remember, proceed from left to right.

It Is Time To Learn Some Greek

Many of the mathematical procedures that you will see in this text involve adding numbers together and then manipulating the total in some manner. For instance, we could add all of the heights of the members of a softball team and then divide by the number of players. This would give us what you probably learned is called the average of the heights. In statistics, this is called the **mean** of the heights. It is the same procedure, just with a different name. Unfortunately, describing how to calculate a mean takes a lot of words. In order to keep the length of this text tolerable, we need to agree upon some simple definitions. Instead of "add all of the heights" or "add all of the scores" we will write $\sum X$. The symbol $\sum$ (the Greek letter capital sigma) indicates that we are to add all the examples of something, and the X stands for a single score or datum (datum is the singular of data, which is always plural). Thus, $\sum X$ says the same thing as "sum each of the scores", but it takes less space. It is read as 'sum of X'. And the mean can thus be written as $\sum X / N$, where N is the total number of scores. Similarly, $\sum X^2$ says the same thing as "sum each of the squared scores". It is read as 'sum of X squared'. But be careful. In mathematics the precise symbols, and the order in which they occur, are important. Just as "Susan, please give the paper to Howard" does not mean the same thing as "Howard, please give the paper to Susan", $\sum X^2$ is not equivalent to $(\sum X)^2$. In the case of $\sum X^2$, we are being told to sum all of the X^2 values. In other words, we first square each score and then add the resulting numbers together (remember, exponents before addition). This is illustrated with a set of numbers given in Table 1.2. Note that $\sum X^2$ is equal to 110. On the other hand, the expression $(\sum X)^2$, which is read as 'sum of X, quantity squared', indicates that we are first to sum all of the X scores and then square the result (remember, operations within parentheses before exponents). In our case, this would be 18^2, which equals 324. Obviously 110 is not the same as 324! It will be important to remember as you read this text to pay careful attention to the details of the mathematical statements. While I have done my best to present concepts clearly, this is not material that can be 'skimmed'.

Mean – Sum of the scores divided by the total number of scores.

Table 1.2 Computing $(\sum X)^2$ and $\sum X^2$

X	X^2
5	25
6	36
7	49
$\sum X = 18$	$\sum X^2 = 110$

$(\sum X)^2 = (18)^2 = 324$

Using Algebraic Expressions

It is also important to point out that you need to be familiar with the manipulations conducted with algebraic equations. For instance, if you are given the rather impressive equation $r_s = 1 - [(6\sum D^2) / n(n^2 - 1)]$ and are told that $\sum D^2$ equals 10, and n equals 5, you need to be able to determine the value of r_s. To do so, simply substitute the value 10 where $\sum D^2$ appears in the numerator of the fraction and substitute the value 5 where n appears in the denominator:

$$r_s = 1 - \frac{6\sum D^2}{n(n^2 - 1)}$$

$$= 1 - \frac{6(10)}{5(5^2 - 1)}$$

$$= 1 - \frac{60}{5(25 - 1)}$$

$= 1 - \frac{\mathbf{60}}{\mathbf{5(24)}}$

$= 1 - \frac{\mathbf{60}}{\mathbf{120}}$

$= 1 - 0.5$

$= 0.5$

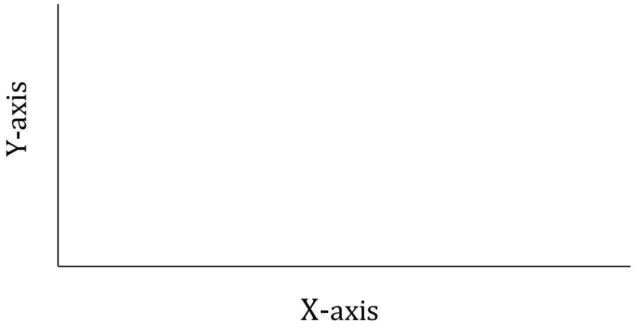
If you successfully followed this calculation the algebraic expressions in this book should not be a problem. By the way, you just calculated your first Spearman correlation, a statistic reviewed in the appendix. Congratulations!

Graphs

You will be seeing many graphs is this book. Graphs are used in statistics because they often simplify complex situations – they allow us to 'see' a relationship more clearly, or at least more easily, than would be the case if expressed only in words or equations.

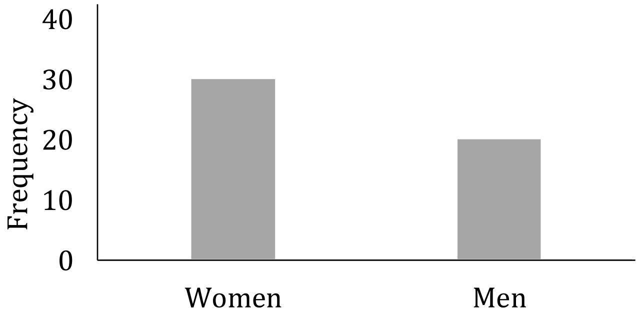
The graphs in this book consist of two lines (called axes), which are labeled X and Y. They are arranged at right angles to each other (Figure 1.2).

Figure 1.2 Basic Form of a Simple Graph



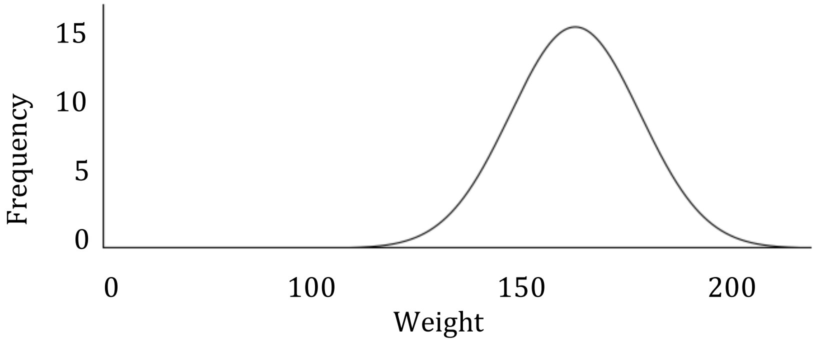
For instance, a simple graph might indicate the number (frequency) of women and men taking a statistics course (Figure 1.3). In this graph the X-axis consists of two categories, women and men, while the Y-axis consists of frequencies. At a glance it is evident from the heights of the two columns that somewhat more women (30) are taking the course than men (20).

Figure 1.3 Frequencies of Women and Men Taking a Course



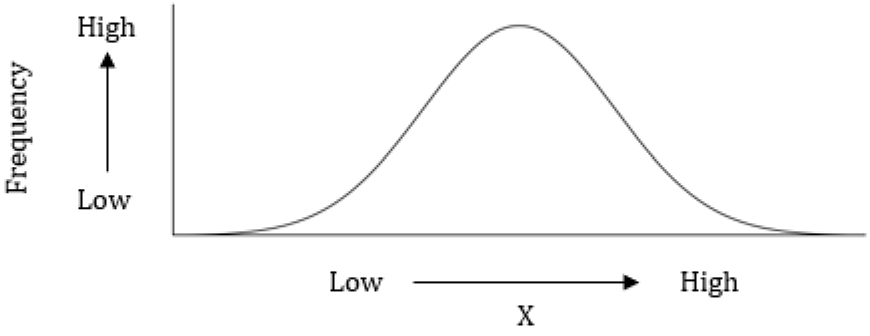
A slightly more complex graph might continue to have frequencies on the Y-axis but now has a series of values on the X-axis. For instance, a graph of the weights of students taking a statistics course might look like Figure 1.4. This graph indicates that most students' weights would be between approximately 120 and 180 pounds, with fewer having lower or higher weights.

Figure 1.4 Weights of Students Taking a Course



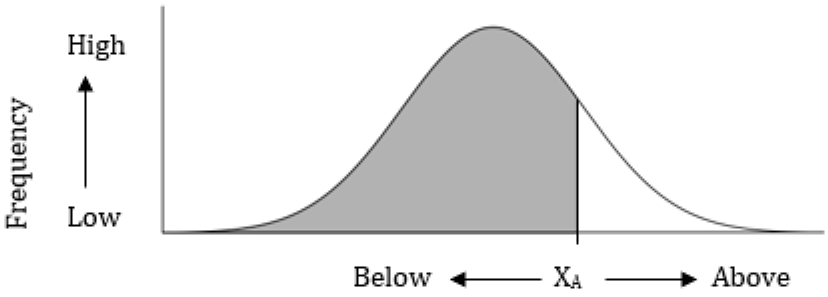
It is important to understand the general organization of this graph as you will be seeing numerous variations throughout this book (Figure 1.5). We have just seen that a variable, weight, was placed on the X-axis. Other variables we might place on the X-axis would include height, IQ scores, exam scores, and grade point average, to name just a few. And in each case we often (but certainly not always) find that most values occur in the middle of the distribution with progressively fewer examples as we move away in either direction. This is illustrated with the curve in Figure 1.5.

Figure 1.5 Commonly Used Organization for a Graph



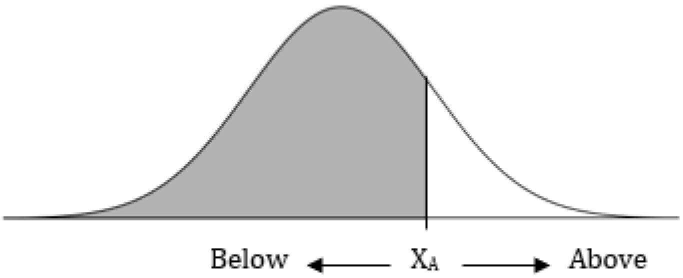
It is also important to understand a couple of additional terms which are associated with graphs. For instance, in Figure 1.6 the region of the curve to the left of a point on the X-axis (I have called it X_A) would be said to be 'below' X_A. This region is shaded in Figure 1.6. Naturally, the region of the curve to the right of this point would be said to be 'above' X_A.

Figure 1.6 Illustration of the Meaning of 'Below' and 'Above'



Finally, please note that in Figures 1.3 and 1.4 specific frequencies were identified on the Y-axis. However, in Figures 1.5 and 1.6 this was not the case. Instead we just had a continuum of frequencies from 'low' to 'high'. In fact, it is so common not to have specific values of the frequencies that the vertical line and label representing the Y-axis are often simply deleted from the graph (Figure 1.7). Additional examples of this occur throughout the text.

Figure 1.7 Reproduction of Figure 1.6 but with the Label for the Y-Axis Omitted



Hopefully this review of graphs has confirmed that they are easy to use. And you will see that graphs can be a great help in illustrating relationships that would otherwise be difficult to describe.

Conclusion

In order to be proficient with statistics you need to understand which statistical procedure to utilize, and why. This book is designed, therefore, to focus upon the ideas that are essential to the understanding of statistical reasoning. You will also need to make accurate calculations but, as you will later see, much of the tedium of number crunching can be eliminated by utilizing a statistical computer package such as SPSS. Finally, you will need to be proficient using graphs.

Glossary Of Terms

Absolute value – The magnitude of a number irrespective of whether it is positive or negative.

Data (plural of datum) – Observations or factual information, often in the form of numbers.

Descriptive statistics – Techniques that are used to summarize data. These procedures lead to a better understanding of the data.

Inferential statistics – Techniques that are used in making decisions based upon data.

Mean – Sum of the scores divided by the total number of scores.

Questions – Chapter 1

(Answers are provided in Appendix J.)

1. Knowledge of probabilities was initially used with _____.
   a. Scientific studies
   b. Voyages of discovery
   c. Games of chance
   d. Voting in elections

2. The most rapid period of change in the field of statistics occurred in the _____ centuries.
   a. Late 17th and early 18th
   b. Late 18th and early 19th
   c. Late 19th and early 20th
   d. Late 20th and early 21st

3. Basing decisions upon limited exposure to the relevant information is known as the _____ bias.
   a. Availability
   b. Probability
   c. Limited exposure
   d. Information

4. What does (–2) / (–4) equal?
   a. .5
   b. 8
   c. –.5
   d. –8

5. What does $2X + Y^2$ equal if X is 3 and Y is 6?
   a. 21
   b. 28
   c. 42
   d. 95

6. What does –2(36) equal?
   a. –72
   b. –18
   c. 34
   d. 118

7. I have a large set of data and wish to present it to an audience in a form that is easy for them to understand. This is an example of _____.
   a. inferential statistics
   b. descriptive statistics

8. What is $\sum X$ and $\sum X^2$ for the set of numbers consisting of 2, 3 and 4?
   a. 9 and 81
   b. 24 and 29
   c. 9 and 29
   d. 9 and 9

9. What is $\sum X$ and $\sum X^2$ for the set of numbers consisting of −2, 3 and 4?
   a. −9 and 81
   b. 5 and 29
   c. 9 and 29
   d. −9 and −9

10. The procedures that are used to describe large amounts of data in quickly understandable ways are called _____.
    a. inferential statistics
    b. descriptive statistics

11. What does (−2) / 4 equal?
    a. 0.5
    b. 8
    c. −0.5
    d. −8

12. What does (−3) – (−5) equal?
    a. −8
    b. −2
    c. 2
    d. 15

13. What does (−6) (−3) equal?
    a. 18
    b. −9
    c. −3
    d. −18

14. What is the positive square root of 144?
    a. 10
    b. 11
    c. 12
    d. 13

15. Which of the following statements is equivalent to $X > 6$ and $Y < 22$?
    a. X is less than 6; Y is greater than 22
    b. X is greater than 6; Y is less than 22
    c. X is less than 6; Y is less than 22
    d. X is greater than 6; Y is greater than 22

16. The equation for the variance of a population (this will be covered in Chapter 3, don't worry about the definition of the symbols) is $\sigma^2 = (\Sigma (X - \mu)^2) / N$. If $(\Sigma (X - \mu)^2$ is equal to 8, and N is equal to 3, what does σ^2 equal?
    a. 2.67
    b. .375
    c. 24

d. 10.26

17. What does (–6) / (–3) equal?
a. –2
b. –18
c. 18
d. 2

18. What does 36 / (–2) equal?
a. –2
b. –18
c. 18
d. 2

19. If the proportion of people who like to eat apples is .42, and there are 2000 people, how many people like to eat apples?
a. 420
b. 840
c. 4200
d. 8400

20. The absolute value of 36 is _____.
a. √36
b. 36^2
c. 36
d. –36

21. The absolute value of –14 is _____.
a. √14
b. 14^2
c. 14
d. –14

22. Assume we asked people to indicate whether they would prefer a cat or a dog as a pet, and then created a graph. The preferences could be indicated with two columns, one for preference for cats, the other for preference for dogs. The height of each column would indicate the _____ of the preference.
a. Value
b. Wisdom
c. Certainty
d. Frequency

23. In a graph, if you are to the right of a particular point on the X-axis we would say you are _____ this point.
a. Above
b. Below
c. Under

DESCRIPTIVE STATISTICS – SEEING YOUR DATA MORE CLEARLY

Chapter 2 – Describing Nominal and Ordinal Data

The Descriptive Statistics Used with Nominal and Ordinal Data

Chapter 3 – Describing Interval and Ratio Data – I

An Introduction to the Descriptive Statistics Used with Interval/Ratio Data

Chapter 4 – Describing Interval and Ratio Data – II

Further Descriptive Statistics Used with Interval/Ratio Data

Chapter 2
Describing Nominal and Ordinal Data: The Descriptive Statistics Used with Nominal and Ordinal Data

"Whenever you can, count."

Sir Francis Galton

All Numbers Are Not Equal

We are all familiar with measuring things. In the United States, the gas you use to fill the tank of your car is measured in gallons. Your height is measured in feet and inches. Your weight is measured in pounds. In most of the world, you would have used liters, centimeters, and kilograms for these measurements. Obviously, we often find that what we measure varies. For instance, cars have gas tanks of different sizes, and heights as well as weights of people vary. When what we are measuring can vary, we call it a **variable**.

Variable – Any characteristic that can vary.

It is likely that you have not given much thought to the implications of how we measure variables. For instance, at the Olympics it does not matter if we know precisely how quickly three runners completed a race. As long as we know their order of finishing we can hand out the medals properly. However, for statistics it matters a great deal. You will learn that whether you simply measure the order, or instead the elapsed durations, of runners has implications for the proper choice of statistical technique to employ.

Sometimes A Number Is Just A Name Or Category

The **nominal scale of measurement** provides the least amount of information. As the word nominal implies, with a variable measured on a nominal scale we are using a number in place of a name, and thus the number serves as a label. Put another way, with a nominal scale of measurement we are using numbers to assign individuals to categories. For instance, we commonly describe a child as being either a boy or a girl. And we could arbitrarily assign the number 1 to each boy and the number 2 to each girl. In this case, the number simply indicates the group to which each individual belongs. The only data that are meaningful would be how many individuals are members of each group. Thus, it might be important for a school to know how many boys and how many girls are enrolled each semester. Notice that it makes no sense to argue that because a boy is a 'one' and a girl is a 'two', a girl is twice what a boy is. In other words, multiplication or division, for instance, cannot be used with these data. Since the labels of the groups were assigned arbitrarily, we could have instead assigned the number 'one' to each girl and the number 'two' to each boy. Alternatively, we could have given each boy a label of 'zero' and each girl a label of 'nine'. It does not matter. With a nominal scale of measurement the number is just a label; the magnitude of the actual label is not meaningful. All that is meaningful is the frequency of individuals in each group.

Nominal scale of measurement – A measurement scale in which numbers serve as names of categories. In this level of measurement, the magnitude of the number is arbitrary.

Sometimes A Number Tells Us The Order Of Events

With the **ordinal scale of measurement** we know the order in which events occurred. Thus we have more information than with a nominal scale. For instance, where a person places at the conclusion of a footrace is an example of ordinal data. With ordinal data we know that whoever came in first had to get to the finish before whoever came in second. What we do not know is how much sooner the first place finisher completed the race compared to the second place finisher. It might have been a photo finish, or there may have been enough time for the first place runner to shower and go home before the runner in second place completed the race. In other words, with ordinal data we know the order of events, but we do not know the magnitude of the difference between events.

Ordinal scale of measurement – A measurement scale in which the magnitude of the numbers indicates the order in which events occurred. In this level of measurement, the magnitude of the number is meaningful.

Sometimes We Can Add And Subtract Numbers

Data on an **interval scale of measurement**, in contrast, not only indicate the order of events but also the magnitude of the difference between events. As a result, interval data provide more information than ordinal data. It is now appropriate, for the first time, to use addition and subtraction. For instance, we can now say not only that one day is colder than another day, we can use subtraction to find how many degrees colder. However, as you will see, with data measured on an interval scale it is still not appropriate to multiply or divide the scores.

The two most popular temperature scales, Fahrenheit and Centigrade, are examples of interval scales of measurement. A characteristic of interval scales is that though they may have a zero point this does not indicate the complete absence of whatever is being measured. Thus, there is a 0 degrees Fahrenheit, but it does not indicate that this is the coldest possible temperature. I live near Buffalo, New York, and can attest to the fact that it may drop below 0 degrees Fahrenheit during the winter. On those occasions the temperature is measured in negative numbers. The same situation occurs for the Centigrade scale. In the Centigrade scale, 0 degrees is defined as the freezing point of water. Your refrigerator's freezer should have a temperature below 0 degrees Centigrade. Since there is no absolute zero point in either the Fahrenheit or Centigrade scales it is not appropriate to multiply or divide these numbers. Thus, a day with a temperature of 90 degrees Fahrenheit is not three times as hot as a day that has a temperature of 30 degrees. However, as it is appropriate to add or subtract with an interval scale of measurement, we can say that the 90 degree day is 60 degrees warmer than the 30 degree day.

<u>Interval scale of measuremen</u>t – A measurement scale in which the magnitude of the difference between numbers is meaningful, and thus addition and subtraction are possible. However, there is no true zero and thus multiplication and division are not meaningful.

Some Numbers Can Be Multiplied And Divided

The last of the four scales is the **ratio scale of measurement**. In addition to having the characteristics of an interval scale, a ratio scale also has an absolute zero point. Those of you who have taken chemistry may recall that with the Kelvin scale there is a true zero point below which it cannot get colder. Therefore, the Kelvin scale is a ratio scale. Time is another example of a variable measured with a ratio scale. A race starts at time zero and proceeds to the finish. Since there is a true zero with time, we can not only subtract runners' times to look at differences, we can also meaningfully multiply and divide their times. Thus, if one runner completes a race in 2 minutes and another finishes in 4 minutes, we can say that the first runner was twice as fast as the second. It is only with data on a ratio scale that we can meaningfully say that one number is a multiple of another.

<u>Ratio scale of measuremen</u>t – A measurement scale in which the magnitude of the difference between numbers is meaningful, and there is a true zero. Thus, multiplication and division as well as addition and subtraction are meaningful.

The Big Picture

We have just learned that there are four measurement scales. From having the least to most information, the order of these scales is: nominal, ordinal, interval and ratio. The data in a scale with more information can be converted into a scale with less information, but not the reverse. For instance, if you use a stopwatch to time a race (ratio scale), you could then compare the times to assign medals (ordinal scale). But if all you record is the order that the runners finished the race (ordinal scale) you cannot subsequently establish each runner's actual time (ratio scale). And you should recognize that whenever data measured on a more informative scale are converted to a less informative scale, information is being lost.

From our perspective, the distinction between scales of measurement is important because choosing the appropriate statistical procedure depends, in part, upon the scale of measurement that was utilized. Thus, the statistical procedures utilized with nominal data differ from those used with ordinal data. And while the same statistical procedures are used with interval and ratio data, these procedures are distinct from those used with either nominal or ordinal data. Also, in Chapter 1 you learned that the two major functions of statistics are to assist you in seeing and thinking about data more clearly. You will recall that these uses are referred to as descriptive and inferential statistics, respectively. This is important because the question you are asking – are you trying to see or think about data more clearly – is also important in determining the appropriate statistical procedure. How the question you are asking (descriptive or inferential statistics) and the measurement scale you utilize interact can easily be illustrated (Table 2.1) and would provide a logical basis for the order of coverage of topics in an ideal statistics course. Unfortunately, time constraints require that some topics be omitted in a brief coverage of statistical procedures. As a result, the order that topics will actually be covered in this text is illustrated in Table 2.2.

Table 2.1 A Logical Order of Coverage of Topics in a Statistics Course

		Type of Data		
		Nominal	Ordinal	Interval/Ratio
Type of Statistics	Descriptive	1	2	3
	Inferential	4	5	6

Table 2.2 Actual Order of Coverage in this Text

		Type of Data		
		Nominal	Ordinal	Interval/Ratio
Type of Statistics	Descriptive	1	2	3
	Inferential	4	Appendix	5

A Further Distinction: Discrete Versus Continuous Data

You have just learned that nominal data refer to categories and ordinal data deal with ordered events. In both cases there can only be particular values. With nominal data, an individual is either included in a category, or not. With ordinal data, a runner comes in first, second, or third, not some intermediate value. When a variable can only have particular values, it is said to be **discrete**. By contrast, some variables, such as distance, can take on any value. You can go 5.0 miles, or 5.2361 miles. There are not limited magnitudes that the variable is restricted to, at least within some overall range of possible values. In these cases, in which the magnitude of the variable is not restricted to particular values, the variable is said to be **continuous**. Most examples of interval and ratio data are continuous. However, in some cases data can only take on particular values and yet are considered to be continuous. For instance, the score you receive on a 100-point multiple choice exam can only consist of whole numbers, such as 85 or 91. You cannot have an intermediate score, such as 74.92. Nevertheless, this would be treated as being a continuous variable as there are numerous possible values.

Discrete variable – A variable that can only have particular values.

Continuous variable – A variable that can be of any magnitude, though it might be limited to a particular range.

"In god we trust. All others must bring data."

Attributed to W. Edwards Deming

Seeing Clearly With Nominal Data: An Introduction To Descriptive Statistics

It's Time To Begin Learning To See Better

We will now begin to learn how to see or understand data more accurately. We will start with nominal data, those data that are used as names or categories. We are, therefore, beginning with position '1' in Table 2.2, the descriptive statistics of nominal data. Remember, these data are discrete, and with nominal data you cannot meaningfully add, subtract, multiply or divide. All you have are the frequencies for each category.

Let's assume that we have asked the members of a college class to identify their political party affiliations. The students are found to be Democrats (D), Libertarians (L), Republicans (R), Socialists (S), or to have no party affiliation (N). The hypothetical results for the 25 students are indicated in Table 2.3.

Table 2.3 Data of Students' Political Party Affiliations

N	R	D	D	N
L	N	N	R	D

R	N	D	S	N
N	D	R	N	D
R	N	L	N	D

In this form, it is not easy to quickly understand the students' party affiliations. About all one can rapidly discern is that most students appear to be Democrats (D), Republicans (R) or they have no party affiliation (N). If the data are organized so that the frequency of each party affiliation is recorded, then we have what is called a **frequency distribution** (Table 2.4, in this example the party affiliations are listed alphabetically). With a frequency distribution it is easy to quickly gain an overview of the data.

Table 2.4 Frequency Distribution of Student Political Party Affiliation

Party Affiliation	Frequency
Democrat	7
Libertarian	2
No Affiliation	10
Republican	5
Socialist	1
Total	25

I think you will agree that simply organizing the scores into a frequency distribution aids in the rapid understanding of the data. In other words, you are beginning to see the data more clearly.

Frequency distribution – A listing of the different values or categories of the observations along with the frequency with which each occurred.

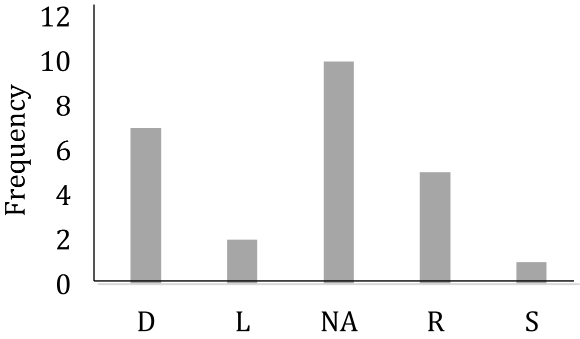
Once a frequency distribution has been constructed it is then easy to determine the **relative frequency** for any category. To do so, we divide the frequency of a category by the total frequency. Referring to Table 2.4, the relative frequency of Republicans would be 5 / 25 which is equivalent to 1 / 5 or 0.20.

Relative frequency – The frequency of a category divided by the total frequency.

Using Graphs And Charts To See Nominal Data More Clearly

For nominal data, two of the most commonly used techniques for summarizing findings are the **bar graph** and the **pie chart**. With a bar graph each category of response is usually identified on the X-axis, and the frequency with which it occurred is usually noted on the Y-axis. As we are dealing with separate categories, in the bar graph the 'bars' representing the frequencies are drawn so that they do not touch each other. A bar graph for our hypothetical political affiliation data is indicated in Figure 2.1.

Figure 2.1 Bar Graph of Student Political Party Affiliation



Note: D, Democrat; L, Libertarian; N, no affiliation; R, Republican; S, Socialist

With a bar graph the reader can quickly 'see' the political party preferences of the students. Because bar graphs are an effective way to present summary data you will commonly see them in newspaper articles as well as in magazines. However, the bar graph is not the only choice for representing a set of nominal data.

<u>Bar graph</u> – A graph in which the frequency of each category or class of observation is indicated by the length of its associated bar.

The pie chart is also sometimes used to summarize nominal data. With a pie chart the frequency of each category of responses is first converted into a relative frequency. As was noted previously, this is accomplished by dividing the frequency for each category of response by the total number of responses (the total of all of the frequencies for all of the categories), in our case 25. For instance, 10 students indicated that they do not have a party affiliation. To find the relative frequency of these 10 students, we would divide 10 by 25, the total number of responses. This would be 0.40, which is equivalent to 40%. The result of the calculations for each category is shown in Table 2.5.

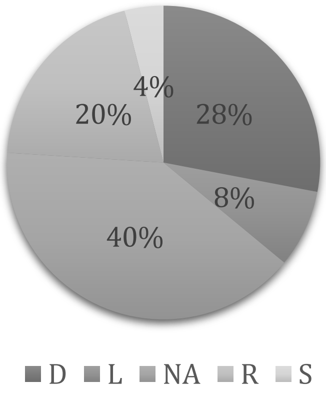
Table 2.5 Frequency Distribution of Student Political Party Affiliation With Associated Relative Frequencies

Party Affiliation	Frequency	Relative Frequency
Democrat	7	.28
Libertarian	2	.08
No Affiliation	10	.40

Republican	5	.20
Socialist	1	.04
Total	25	1.00

Once the relative frequencies are calculated, the area of a circle is divided into slices. There are as many slices as there are categories of response in the frequency distribution, and the area of each slice corresponds to the relative frequency of each category (Figure 2.2).

Figure 2.2 Pie Chart of Student Political Affiliation



D, Democrat; L, Libertarian; NA, No Affiliation; R, Republican; S, Socialist

The pie chart can be an effective way to describe the data. However, its effectiveness is compromised if there are too many categories. Remember, the goal is to convey information efficiently, not to create a visually impressive, but overwhelming, presentation.

Pie chart – A presentation of categorical data in which the area of a slice of a circle is indicative of the relative frequency with which the category occurs.

We have just reviewed how bar graphs and pie charts can be beneficial in summarizing a set of nominal data. In addition to these depictions of the frequency distribution, a reader can also benefit from a single measure that summarizes the entire set of responses. This would be what statisticians call a **measure of central tendency**. An average is an example, so you are familiar with this concept.

Measure of central tendency – A single number that is chosen to best summarize an entire set of numbers.

With nominal data, the **mode** is the most appropriate measure of central tendency. The mode is simply the category with the highest frequency. With our data of student party affiliations, the ‘no affiliation’ response would be the mode, as more of our hypothetical students chose this option than any other. The mode has the advantage that it is very easy to calculate and understand. However, it has a major limitation; it is **unstable**. In many instances if only a few responses were to change, then the mode would change to a different category. Thus, if only two of the students who had indicated that they did not have a party affiliation had, instead, chosen Democrat, then the mode would have shifted to this new category.

Mode – A measure of central tendency. It is the most common category or score.

Unstable – A term used to describe a measure, such as of central tendency, that can vary significantly with only a few changes to the original set of data. This is an undesirable quality.

If one category has the highest frequency, we have a **unimodal** distribution. It is also possible that two, or more, categories will have the same highest frequency. If two categories are tied for the highest frequency, both categories would be modes and the distribution would be said to be **bimodal**. If three categories were tied for the highest frequency, the distribution would be said to be trimodal, and so on.

Unimodal – A descriptive term for a distribution that has one mode.

Bimodal – A descriptive term for a distribution that has two modes.

There is no adequate measure of **variability** for use with nominal data. Variability refers to how much scores differ or deviate from each other. The closest you could come with nominal data would be to simply indicate how many response categories the subjects in a sample had either chosen or been assigned to. With our example, students identified five political party affiliations.

Variability – How much scores differ or deviate from each other.

Summary Of The Descriptive Statistics Of Nominal Data

From this brief review it should be evident that there is nothing particularly challenging about the descriptive statistics of nominal data. Once a frequency distribution is constructed, a bar graph or pie chart and the mode(s) are easy to obtain.

The following table will clarify what you have learned so far in this chapter (Table 2.6).

Table 2.6 Descriptive Statistics of Nominal Data

	Type of Data
	Nominal (Frequency)
Descriptive Procedures (Summarizing the Data)	

Frequency Distribution	Bar Graph or Pie Chart
Central Tendency	Mode
Variability	– – – – –

Progress Check

1. In this measurement scale all that is meaningful is the frequency of events or individuals in each category.
2. In this measurement scale there is a zero, but it is not a true, or absolute zero.
3. When you list the frequency associated with each value or category, you have created a ____.

Answers: 1. Nominal 2. Interval 3. Frequency distribution

Seeing Clearly With Ordinal Data: Continuing With Descriptive Statistics

With ordinal data you are able to rank a set of data along some dimension. For instance, you might know the order in which runners finished a race, or it might be possible to rank 25 students from most to least outgoing. Thus you know who is first, second, and so on, but with data in this form there is nothing further that can be done to increase a reader's understanding. All that is known, or knowable, is the ranking of the runners or that a total of 25 students were ranked on how outgoing they were.

However, if instead of being ranked individually each of the 25 students was assigned to one of five ranked levels, from 'very shy' to 'very outgoing', we could then create a frequency distribution as shown in Table 2.7.

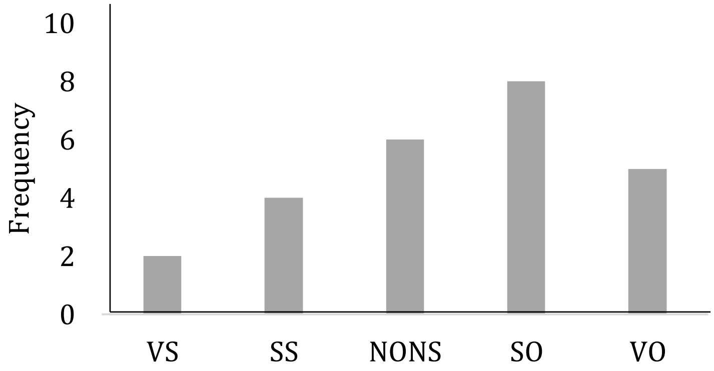
Table 2.7 Frequency Distribution of Being Outgoing

Response Category	Frequency
Very Outgoing	5
Somewhat Outgoing	8
Neither Outgoing nor Shy	6

Somewhat Shy	4
Very Shy	2
Total	25

This frequency distribution allows us to 'see' the data more clearly. And, just as was the case with nominal data, with ordinal data it is simple to create a bar graph once a frequency distribution has been made. However, while the order in which the categories are presented is arbitrary with nominal data, the order is meaningful with ordinal data. Therefore, when ordinal data are assigned to ranked categories, in our case from 'very shy' to 'very outgoing', the bar graph should be organized to reflect this order, as is shown in Figure 2.3.

Figure 2.3 Bar Graph of Ratings of How Outgoing Students Are



Note: VS, Very Shy; SS, Somewhat Shy; NONS, Neither Outgoing Nor Shy;

SO, Somewhat Outgoing; VO, Very Outgoing

Clearly, a properly constructed bar graph permits a rapid understanding of the data.

With ordinal data a pie chart is not appropriate. The problem is that the categories would wrap around the circle so that the two most extreme categories would end up side by side. Thus, the 'very shy' category would be next to the 'very outgoing' category, which would make it more difficult to recognize the order in the responses.

As ordinal data involve ranks, it is now possible to specify relative standings. The **percentile rank** is the percentage of the distribution within or below a category. For instance, referring to Table 2.7 will indicate that 20 of the 25 students, or 80%, indicated that they were either 'somewhat outgoing' or less than 'somewhat outgoing'. Thus a student who was somewhat outgoing would be at the 80th percentile. And the percentile rank for 'neither outgoing nor shy' would be almost 50% (12 / 25 = 48%).

Percentile rank – The percentage of the data at or below a category or score.

The median is the measure of central tendency employed with ordinal data. The **median** can be defined in a number of ways. Perhaps the simplest is that the median is the value that has as many scores above it as below it. In other words, it is the value that divides the distribution into two equal parts. It follows that this value will have a percentile rank of 50% and thus is at the 50th percentile. In our example with 25 subjects, the median would be the value associated with the 13th individual from either end of the distribution. This individual is at the midpoint of the distribution, with 12 entries above, and 12 below. From our frequency distribution, it is evident that this individual would be in the 'somewhat outgoing' category.

If the distribution had an even number of ranks the procedure for finding the median is slightly more involved. For instance, if the distribution consisted of four ranks, 2, 4, 6, and 9, there is no rank at the midpoint of the distribution. In such a situation, the median would be the value halfway between the two mid-most ranks. These ranks are 4 and 6 and the rank halfway between them is 5. (Calculation of the median when used with interval or ratio data is discussed in Chapter 3.)

Median – A measure of central tendency. It is the mid-most score in a distribution. In other words, the median splits a distribution in half, with just as many scores above it as below it. It is at the 50th percentile.

Unlike the case with nominal data, with ordinal data there are measures of variability. The simplest is the **range**. The range is based on the two most extreme data points. For the data provided in Table 2.7, each student was assigned to one of five categories, with the range being from 'very shy' to 'very outgoing'. If there was another set of data which consisted of the numerical ranks in an athletic contest, the range would be determined by finding the difference between the lowest and highest ranks. Thus, if a high school swimming team obtained the ranks of 2, 4, and 12 in a race, the range would be 12 – 2 = 10. Two additional measures of variability that are also sometimes used with ordinal data, the interquartile range and the semi-interquartile range, will be discussed in the next chapter.

Range – A measure of variability for ordinal data. It is obtained by subtracting the lowest rank from the highest rank.

Conclusion

Table 2.8 reviews the descriptive statistics used with nominal data and ordinal data.

Table 2.8 Descriptive Statistics of Nominal and Ordinal Data

	Type of Data	
	Nominal (Frequency)	Ordinal (Ranked)
Descriptive Procedures (Summarizing the Data)		
Frequency Distribution	Bar Graph or Pie Chart	Bar Graph
Central Tendency	Mode	Median
Variability	– – – – –	Range

Hopefully you will agree once again that there is nothing challenging about these descriptive statistics. Once a frequency distribution is made, appropriate measures of central tendency and variability (for ordinal data) are easy to obtain.

Glossary Of Terms

Bar graph – A graph in which the frequency of each category or class of observation is indicated by the length of its associated bar.

Bimodal – A descriptive term for a distribution that has two modes.

Continuous variable – A variable that can be of any magnitude, though it might be limited to a particular range.

Discrete variable – A variable that can only have particular values.

Frequency distribution – A listing of the different values or categories of the observations along with the frequency with which each occurred.

Interval scale of measurement – A measurement scale in which the magnitude of the difference between numbers is meaningful, and thus addition and subtraction are possible. However, there is no true zero and thus multiplication and division are not meaningful.

Measure of central tendency – A single number that is chosen to best summarize an entire set of numbers.

Median – A measure of central tendency. It is the mid-most score in a distribution. In other words, the median splits a distribution in half, with just as many scores above it as below it. It is at the 50th percentile.

Mode – A measure of central tendency. It is the most common category or score.

Nominal scale of measurement – A measurement scale in which numbers serve as names of categories. In this level of measurement, the magnitude of the number is arbitrary.

Ordinal scale of measurement – A measurement scale in which the magnitude of the numbers indicates the order in which events occurred. In this level of measurement, the magnitude of the number is meaningful.

Percentile rank – The percentage of the data at or below a category or score.

Pie chart – A presentation of categorical data in which the area of a slice of a circle is indicative of the relative frequency with which the category occurs.

Range – A measure of variability for ordinal data. It is obtained by subtracting the lowest rank from the highest rank.

Ratio scale of measurement – A measurement scale in which the magnitude of the difference between numbers is meaningful, and there is a true zero. Thus, multiplication and division as well as addition and subtraction are meaningful.

Relative frequency – The frequency of a category divided by the total frequency.

Unimodal – A descriptive term for a distribution that has one mode.

Unstable – A term used to describe a measure, such as of central tendency, that can vary significantly with only a few changes to the original set of data. This is an undesirable quality.

Variability – How much scores differ or deviate from each other.

Questions – Chapter 2

(Answers are provided in Appendix J.)

1. The number of correct answers on an exam with 50 items would be an example of which scale of measurement?
   a. nominal
   b. ordinal
   c. interval
   d. ratio

2. In a history course you learn that World War II began in 1939. The year is an example of which scale of measurement?
   a. nominal
   b. ordinal
   c. interval
   d. ratio

3. Over a summer, a tourist travels 3,000 miles visiting national parks in the Western United States. Miles are an example of which scale of measurement?
   a. nominal
   b. ordinal
   c. interval

d. ratio

4. At a car show awards are given for the best, second best and third best automobiles. This is an example of which scale of measurement?
a. nominal
b. ordinal
c. interval
d. ratio

5. A bar graph is used with _____ and _____ data.
a. nominal and ordinal
b. ordinal and interval
c. interval and ratio
d. nominal and ratio

6. A pie chart is used with _____ data.
a. nominal
b. ordinal
c. interval
d. ratio

7. For ordinal data the _____ is the measure of central tendency.
a. mean
b. median
c. mode

8. If we graphed the heights of a large group of men and women, we might expect to find a distribution with two peaks, one corresponding to the most frequent height of men and the other corresponding to the most frequent height of women. This would be an example of a _____ distribution.
a. Unimodal
b. Bimodal
c. Trimodal

9. In addition to giving a measure of central tendency, such as the median, a measure of how much a set of scores differ is also commonly provided. This second piece of information is called a measure of _____.
a. variability
b. indecisiveness
c. incompleteness

10. Do nominal data have an adequate measure of variability?
a. Yes
b. No

11. The measure of central tendency for nominal data is the _____.
a. Mean
b. Median
c. Mode

12 This is the only scale in which multiplication and division are meaningful.
a. nominal
b. ordinal
c. interval

d. ratio

13 The only information provided with nominal data is _____.

a. the frequency of events within each category
b. the order that events occurred, such as in an athletic competition
c. greater than, or less than, but not by how much

14. A measure of variability for ordinal data is the _____.

a. range
b. mode
c. median
d. there isn't a measure of variability for ordinal data.

15. A measure, such as the range, which can vary substantially when only a few scores' values change is said to be _____.

a. preferable to a measure which doesn't vary substantially
b. never to be used
c. unstable

16. The median of the ranks 1, 4, 5, 6, and 17 is _____.

a. 4
b. 5
c. 5.5
d. 6

17. The range of the ranks in question 16 is _____.

a. 4
b. 5
c. 16
d. 18

18. The median of the ranks 4, 5, 6, and 17 is _____.

a. 4
b. 5
c. 5.5
d. 6

19. The range of the ranks in question 18 is _____.

a. 3
b. 4
c. 13
d. 17

20. Which measurement scale provides the <u>least</u> information?

a. Nominal
b. Ordinal
c. Interval
d. Ratio

Chapter 3
Describing Interval And Ratio Data – I:
An Introduction To The Descriptive Statistics Used With Interval And Ratio Data

"Statistical thinking will one day be as necessary for efficient citizenship
as the ability to read and write."
H. G. Wells

Introduction

Chapter 2 ended with a review of the descriptive statistical procedures used with ordinal data. Recall that these data are discrete. We now turn to a discussion of interval and ratio data. These data can be discrete or continuous. For instance, the number of runs scored in a baseball game would be an example of a discrete variable. You score 0, 1, 2 etc. runs. You cannot score 2.3 runs. By contrast, height and weight are examples of continuous variables as intermediate values are possible.

A useful way to gain an overview of a set of interval or ratio data is with what is called the **stem-and-leaf display**. For example, let's assume we list the scores of 24 students on an exam (Table 3.1). This is an example of ratio data as the order as well as the magnitude of the difference between scores is known, and there is a true zero (a student could have gotten no answers correct).

Table 3.1 Twenty four Scores on an Exam

92, 76, 83, 88, 67, 94, 83, 74, 70, 64, 42, 81, 83, 90, 75, 87, 77, 82, 97, 46, 85, 71, 63, 79

Stem-and-leaf display – A commonly used summary of interval or ratio data in which each original score is separated into two parts, a stem and a leaf.

In this form it is not easy to immediately gain an understanding of the entire set of data. A first step would be to arrange all of the scores in ascending order. The lowest score would be 42, and we would proceed until we reached 97, which is the highest score. While helpful, this would still result in a long row of numbers that is difficult to entirely grasp (Table 3.2).

Table 3.2 Twenty four Scores on an Exam in Ascending Order

42, 46, 63, 64, 67, 70, 71, 74, 75, 76, 77, 79, 81, 82, 83, 83, 83, 85, 87, 88, 90, 92, 94, 97

With a stem-and-leaf display we enhance the presentation by separating each number into two parts, the last digit(s), called a **leaf** (in our example this would be the digit in the ones position) and the preceding digit(s) called a **stem** (in our case the number in the 10s position). For our example there would only be five stem values, 4, 6, 7, 8 and 9. These would be arranged in a column (Table 3.3).

Table 3.3 Stem Values of the Original Data

9
8
7
6
4

Leaf – The last digit(s) of a score. With a stem-and-leaf display each leaf is paired with the appropriate stem value and the leaves are listed in ascending order in each row of the display.

Stem – With a stem-and-leaf display, a list of the different values of the data once the last digit(s) of each score is removed.

It is preferable to keep the intervals equal, so Table 3.3 could be improved by including a stem value of 5 even though there were no scores with this stem (Table 3.4).

Table 3.4 Complete List of Stem Values

9
8
7
6
5
4

Of course, a great deal of information has been lost by providing only the stem values. However, if the values of the digits in Table 3.2 that were dropped were now included (remember each of these values is called a leaf) then none of the original information would have been lost (Table 3.5). For example, the first row, which begins with the stem value of 9, consists of the leaves of 0, 2, 4, and 7. These correspond to the original scores of 90, 92, 94, and 97.

Table 3.5 Stem and Leaves

Stem	Leaf

9	0, 2, 4, 7
8	1, 2, 3, 3, 3, 5, 7, 8
7	0, 1, 4, 5, 6, 7, 9
6	3, 4, 7
5	
4	2, 6

In this form a great deal of information can be understood quickly. For instance, it is obvious that most students scored in the 70s and 80s, that three students had scores of 83, and that no students scored between 50 and 59.

Clearly, data can be summarized very effectively with a the stem-and-leaf display. However, there are additional techniques that are commonly used for summarizing interval and ratio data. For instance, let's assume that Table 3.6 lists the hypothetical incomes of 10 students in a college statistics class, rounded to the nearest thousand dollars. Once again, this is an example of data measured at the ratio level as the order as well as the magnitude of the difference between scores is known, and there is a true zero (a student could have earned nothing).

Table 3.6 Income of 10 College Students

Student	Income in Dollars
1	20,000
2	3,000
3	1,000
4	2,000
5	4,000
6	3,000
7	10,000
8	3,000
9	4,000
10	7,000

In this form, the data are hard to understand. However, for greater clarity they can be rearranged in descending order (Table 3.7) or, even better, also converted into a frequency distribution, as shown in Table 3.8.

Table 3.7 Income of 10 College Students in Descending Order

Student	Income in Dollars
1	20,000

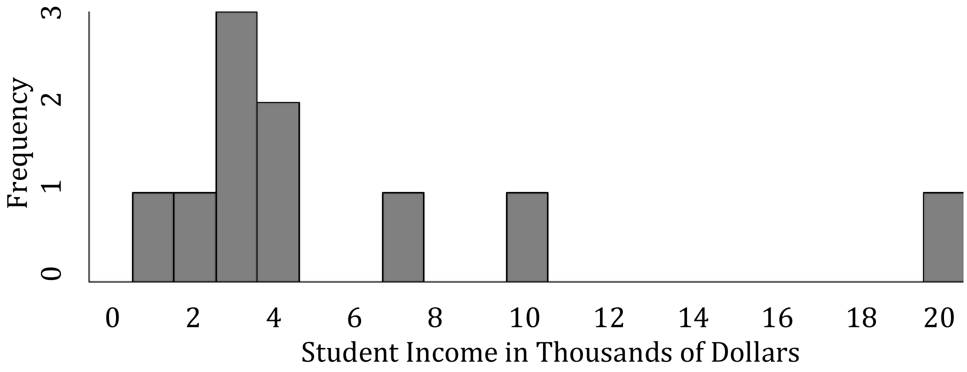
7	10,000
10	7,000
5	4,000
9	4,000
2	3,000
6	3,000
8	3,000
4	2,000
3	1,000

Table 3.8 Frequency Distribution for Student Income Data

Income	Frequency
20,000	1
10,000	1
7,000	1
4,000	2
3,000	3
2,000	1
1,000	1

Once a frequency distribution has been created it is easy to graph the data. With interval or ratio data that are continuous, the graph we would use is either a **histogram** or a **frequency polygon**. A histogram for the data in Table 3.8 is shown in Figure 3.1. Clearly, a histogram looks very much like a bar graph. On both a bar graph and a histogram the values of the responses are usually depicted on the X-axis and the frequencies of the responses are on the Y-axis. However, there are some important differences between the two types of graphs. First, the vertical 'bars' in a bar graph are separated, while the vertical bars in a histogram are positioned side-by-side so that they touch. Further, on the X-axis of a bar graph there are distinct categories, while the intervals on the X-axis of a histogram are specified by what are called **real limits**. The income labeled $2,000, for example, has the lower real limit of $1,500 and the upper real limit of $2,500 because we rounded off to the nearest one thousand dollars and any value from $1,500 to $2,500 was included in the $2,000 category. When there is a score that has the same value as a real limit, it should be randomly assigned to one of the two intervals associated with that limit. In our example, an income of exactly $2,500 could be included in the interval from $1,500 to $2,500, or the interval from $2,500 to $3,500, depending on the flip of a coin. Finally, for some intervals, such as from $8,500 to $9,500, there were no student incomes and thus there is no vertical 'bar'.

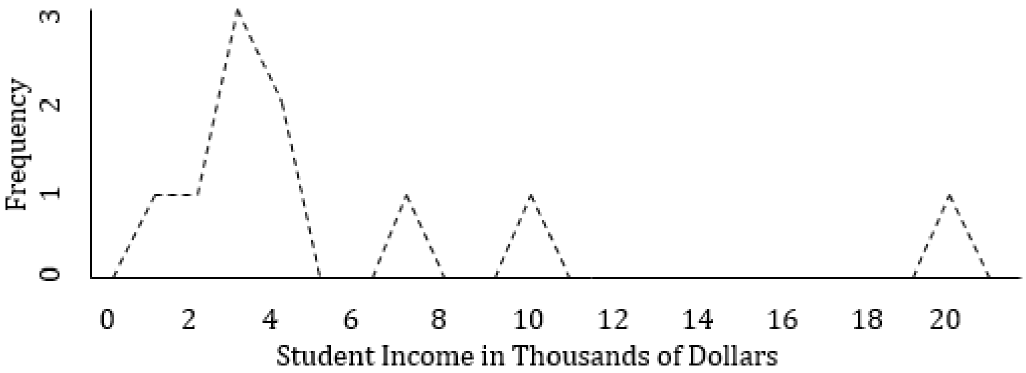
**Figure 3.1	Histogram of the Student Income Data**



<u>Histogram</u> – A graph used with interval/ratio data. As with the bar graph, frequencies are indicated by the length of the associated bars. However, as the data are continuous in a histogram the bars are positioned side-by-side.

It also would be appropriate for these data to be graphed with a frequency polygon. A frequency polygon of the data in Table 3.8 is shown in Figure 3.2. As with a histogram, a frequency polygon is quite easy to construct once a frequency distribution has been constructed. Though a frequency polygon looks somewhat different than a histogram, they are actually closely related. In fact, a frequency polygon can be constructed by simply connecting the center points of each of the vertical 'bars' in a histogram, as is shown in Figure 3.3. With a set of data that has a large number of possible X values a frequency polygon will be easier to construct and read than a histogram.

**Figure 3.2	Frequency Polygon of the Student Income Data**

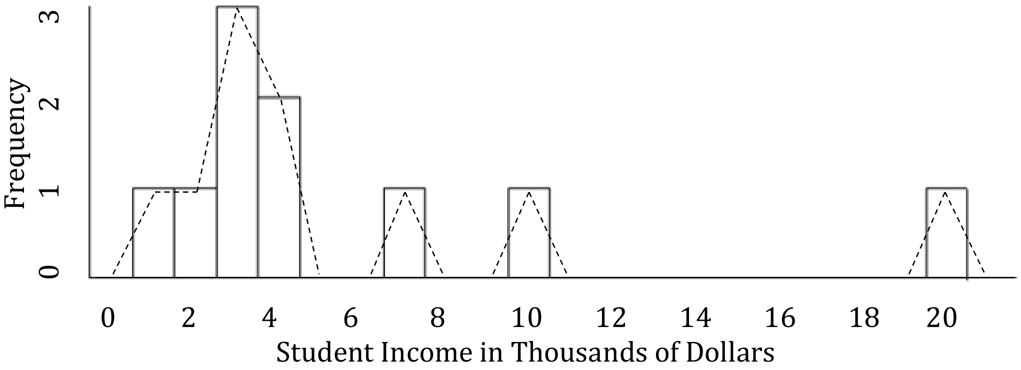


<u>Frequency polygon</u> – A graphic presentation for use with interval or ratio data. It is similar to a histogram except that the frequency is indicated by the height of a point rather

than the height of a bar. The points are connected by straight lines.

<u>Real limits</u> – *With interval or ratio data, the actual limits used in assigning a measurement. These are halfway between adjacent scores, and are called the upper and lower real limits.*

Figure 3.3 Comparison of a Histogram and a Frequency Polygon



Summary To This Point

Thus far in this chapter we have learned how a stem-and-leaf display can provide a useful summary of interval and ratio data, and we have reviewed the advantages of creating a histogram or frequency polygon. We will now turn to a discussion of measures of central tendency.

Measures Of Central Tendency And How The Shape Of The Distribution Affects Their Choice

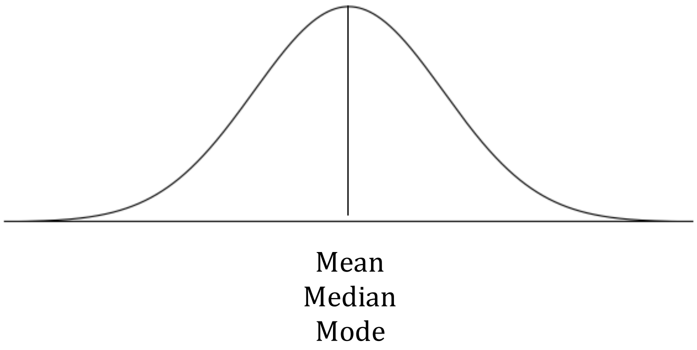
We have seen that with interval or ratio data constructing a stem-and-leaf display and either a histogram or a frequency polygon are straightforward and useful ways to summarize a set of numbers. Calculating a measure of central tendency is also easy. *The mean is generally the preferred measure of central tendency when there are interval or ratio data*. Recall that the **mean** is what most people call an average. To calculate a mean we add all the scores and then divide by the total number of scores. This is symbolized by (ΣX) / N, where N equals the total number of scores. The mean for our 10 student incomes would be $57,000 / 10, which equals $5,700.

<u>Mean</u> – *A measure of central tendency for use with interval or ratio data. It is what is commonly called an average. The mean is the sum of the scores divided by the number of scores.*

The mean is not as easy to conceptualize as the mode or the median. The mean is the balance point of a distribution. In other words, if you made a copy of the frequency polygon in Figure 3.2 out of metal or wood, the point along the X-axis where it would balance would be the mean. The mean is the most frequently used measure of central tendency for interval and ratio data. It has the major advantage that it is used in further statistical procedures that you will learn in later chapters. One unfortunate characteristic, however, is that the mean can be greatly affected by extreme scores. In our set of income data, for instance, the $20,000 response is $10,000 higher than the next highest income. Removing this one income would have a dramatic effect upon the mean. The mean of all ten incomes was $5,700. Without the single $20,000 income, the mean would be $37,000 / 9, or only $4,111. Thus, in this case removing one extreme score results in the mean dropping by over $1,500, or about 28%.

This limitation of the mean can be further understood by looking at either the histogram (Figure 3.1) or the frequency polygon (Figure 3.2). The mean does not appear to be a particularly good single measure of these data. Most of the scores are grouped around $3,000 and $4,000, not around the mean value of $5,700. This is due to the extreme score of $20,000 pulling the mean toward a higher value. This effect of an extreme score will happen whenever the frequency polygon is not symmetrical. In a **symmetrical distribution**, the right half of the distribution is the mirror image of the left half. Figure 3.4 is an example of a symmetrical distribution. In a symmetrical distribution there is a low score that balances the effect that each high score has on the mean.

Figure 3.4 Graph of a Symmetrical Distribution



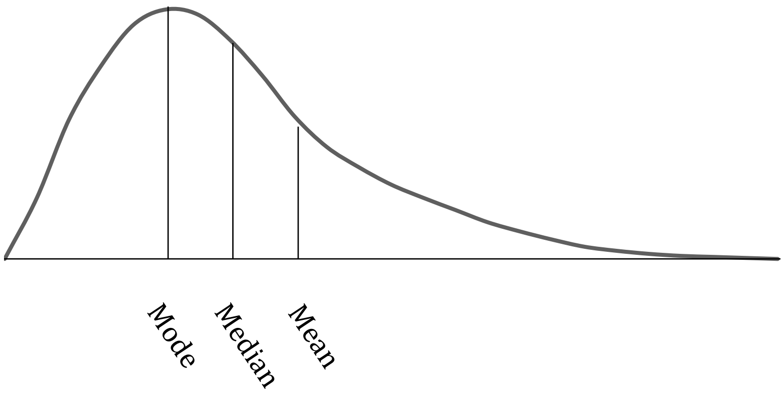
<u>Symmetrical distribution</u> – A distribution in which the right half is the mirror image of the left half. In such a distribution, there is a high score corresponding to each low score.

The curve depicted in Figure 3.4 is a special type of symmetrical distribution referred to as a **bell-shaped curve**. The frequencies are high near the middle, and scores become progressively less frequent the farther they are from the middle. One of the characteristics of a bell-shaped distribution is that the balance point (mean), the middle score (median) and the most frequent score (mode) all have the same value.

<u>Bell-shaped curve</u> – A symmetrical distribution in which the highest frequency scores are located near the middle and the frequency drops the farther a score is from the middle.

The data in Table 3.8 do not form a symmetrical distribution, and are thus said to be **skewed**. More specifically, the distribution appears to look more like Figure 3.5, a nonsymmetrical distribution that points to the right. Such a distribution is called **positively skewed**. The word 'positive' in this context does not indicate 'good', just as the 'positive' terminal of a battery is not 'good'. In both cases, 'positive' is being used to identify an option. In the case of a battery, it is a particular electrical charge. In the case of a graph, it is a direction, the direction of the higher or more positive numbers on a number line.

Figure 3.5 Graph of a Positively Skewed Distribution



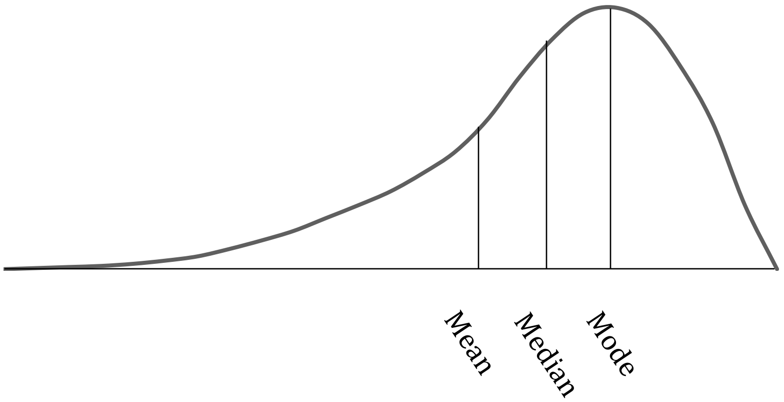
<u>Skewed</u> – A distribution in which one tail is larger than the other. As a result, the distribution is not symmetrical.

<u>Positively skewed</u> – A nonsymmetrical distribution in which the tail pointing to the right is larger than the tail pointing to the left.

In a positively skewed distribution the mean, median and mode do not all fall at the same point. Instead, there is characteristic pattern, as indicated in Figure 3.5. The mode is at the point on the X-axis where the frequency is greatest, the mean is 'pulled' to the right by the extreme scores and the median is located between the mode and the mean. The income distribution in America is an example of a positively skewed distribution. Many people have modest incomes while a few have very large incomes.

It is also possible for a distribution to be **negatively skewed**, as is shown in Figure 3.6. In a negatively skewed distribution the larger tail is pointing toward lower or more negative numbers on a number line. Once again, the mean is being 'pulled' by the extreme scores, except this time to the left; the mode is the value with the highest frequency; and the median is between the mean and the mode. The distribution of scores on an easy exam is an example of a negatively skewed distribution. Many students will do very well, but a few still find the exam to be difficult.

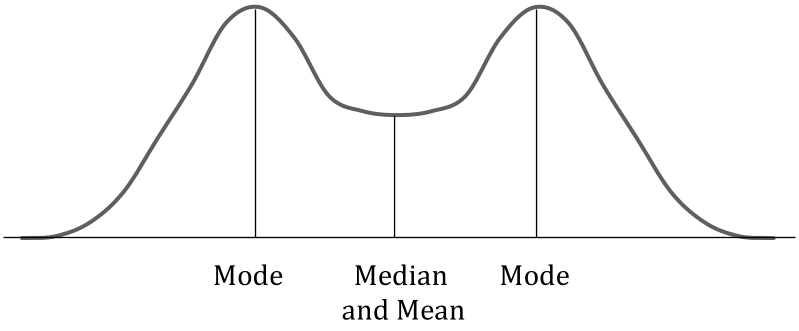
Figure 3.6 Graph of a Negatively Skewed Distribution



Negatively skewed – A nonsymmetrical distribution in which the tail pointing to the left is larger than the tail pointing to the right.

The distributions that have been reviewed thus far are all unimodal, in other words they have only one mode. A symmetrical, bimodal distribution is depicted in Figure 3.7. In a symmetrical, bi-modal distribution, there are two modes, and the mean and the median are located at the same point between these modes. A distribution of heights might be an example of a bimodal distribution, with one mode indicating the most common height for women and the other mode indicating the most common height for men.

Figure 3.7 Graph of a Symmetrical, Bimodal Distribution



We have seen that with interval or ratio data the graph of the frequency distribution can take a number of forms. Whether it is symmetrical or skewed is important, for it affects our choice of statistical procedure. Some statistical procedures assume that the data form a specific bell-shaped curve called a **normal distribution**. With interval or ratio data that are normally distributed the mean is the optimal measure of central tendency. We will see later in the book that the mean has the advantage that it can be used in a variety of flexible statistical procedures. However, when the distribution is not normal, but instead is skewed, we have seen that the mean is 'pulled' by the extreme scores. In that case, the mean would be a poor choice as the measure of central tendency. The median, defined as the midmost score, is less affected by extreme scores and would be a better choice with skewed data. For example, the data presented in Tables 3.7 and 3.8, and graphed in Figure 3.1, are skewed. This is evident as there is a distinctive tail pointing to the right. With these data, even though they were collected at the ratio level, you would probably want to calculate a median rather than a mean as a measure of central tendency.

Normal distribution – *A specific, bell-shaped distribution. Many statistical procedures assume that the data are distributed normally.*

Calculation of the median with interval or ratio data is straightforward:

$$\text{Median} = \text{the value of the score at the } \frac{\mathbf{N+1}}{\mathbf{2}} \text{ position.}$$

If a distribution has an odd number of entries, this equation will result in the median being the middle number in the distribution. For instance, if there were income data from 9 workers, the median would be equal to the income of the worker in the (9 + 1) / 2 position. This would be 10 / 2, which equals 5. In other words, the median would be the income of the fifth worker from the bottom, or top, of the distribution.

The situation is somewhat more complex if there is an even number of entries. With our data there were incomes from 10 students. The median would be the value associated with the (10 + 1) / 2 position on the frequency distribution. This equals 11 / 2, or 5.5. Obviously, there is no 5.5th position. However, we proceed as if there were. The 5th lowest income was $3,000. The 6th lowest income was $4,000. We calculate the mean of these two incomes to find the income of the 5.5th position. In our case this would be ($3,000 + $4,000) / 2. Thus the *median* income of the 10 student is $3,500. It is important to note that the value of this median differs substantially from the value we previously calculated for the mean of the entire set of data, which was $5,700. This difference is due to the distribution being skewed and, consequently, the mean being affected by an extreme income(s). Finally, recognize that in this example the median value of $3,500 is more representative of the incomes of the 10 students than is the value of the mean.

Summary To This Point

I think you will agree that this introduction to the descriptive statistics of interval and ratio data is not any more challenging than the descriptive statistics of nominal or ordinal data. Constructing a stem-and-leaf display is not difficult, and once a frequency distribution is constructed then the histogram or frequency polygon can be constructed easily. And finding the mean is straightforward. Finally, if the data are clearly skewed we discussed that the median would be a better choice as a measure of central tendency than the mean since it is less affected by extreme scores.

Reviewing Table 3.9 may make what you have learned in this chapter clearer as it compares the descriptive statistics used with nominal and ordinal data, which were reviewed in Chapter 2, with the descriptive statistics we have just reviewed for interval and ratio data (underlined in the table). The italicized items in Table 3.9 are additional concepts that will be reviewed in this chapter.

Table 3.9 Overview of Descriptive Statistics

	Type of Data			
	Nominal (Frequency)	Ordinal (Ranked)	Interval/Ratio (Score)	
Frequency Dist	Bar Graph or Pie Chart	Bar Graph	Histogram or Frequency Polygon	
Central Tendency	Mode	Median	IF NOT NORMAL Median	IF NORMAL Mean (Median – less common)
Variability	– – – –	Range	*Interquartile Range*	Standard Deviation z Score[a]
Summary Presentation			Stem-and-leaf display and *Boxplot*	Stem-and-leaf display and *Boxplot*

[a] This procedure is reviewed in Chapter 4

Progress Check

1. To graph the frequency distribution of interval or ratio data we would use either a ____ or a ____.
2. If we measure the duration of a foot race in seconds, then the real limits for a time of 36 seconds would be _____ and _____ seconds.
3. In most soccer games even the winning team ends with a modest score. However, there are exceptions where the winning team has a high score. This is an example of a _____ skewed distribution.

Answers: 1. Histogram; frequency polygon 2. 35.5; 36.5 3. Positively

Measuring The Variability Of Interval And Ratio Data

"Then there is the man who drowned crossing a stream with an average depth of six inches."

W. I. E. Gates

We next turn to another way to provide a summary presentation of interval and ratio data. This is the boxplot, which is based upon the interquartile range, a commonly used measure of variability for interval and ratio data.

Following the discussion of the boxplot, and continuing in the next chapter, we will be focusing on two additional measures of variability that are commonly used with normally distributed, interval and ratio data. These are the standard deviation and the z score. Each is based upon the mean. This chapter will introduce the standard deviation; Chapter 4 will describe the z score.

The Boxplot

A straightforward measure of variability used with interval and ratio data is the **range**. We have briefly discussed the range previously. It is simply the spread of the scores. More specifically, *with interval or ratio data the range is commonly defined as the difference between the highest and lowest scores.** In the case of our income data (Table 3.8), the highest value that was reported was $20,000, and the lowest recorded income was $1,000. The range of the incomes is, therefore, $20,000 – $1,000, which is $19,000. That is all that is involved with calculating the range.

<u>*Range*</u> *– A measure of variability. It is commonly defined as the value which is obtained*

*when the lowest score is subtracted from the highest score.**

**More precisely, the range is defined as the interval that includes all of the scores. Thus, with interval or ratio data it actually equals the difference between the upper real limit of the highest score or category and the lower real limit of the lowest score or category. However, this more precise definition is rarely used in the social sciences.*

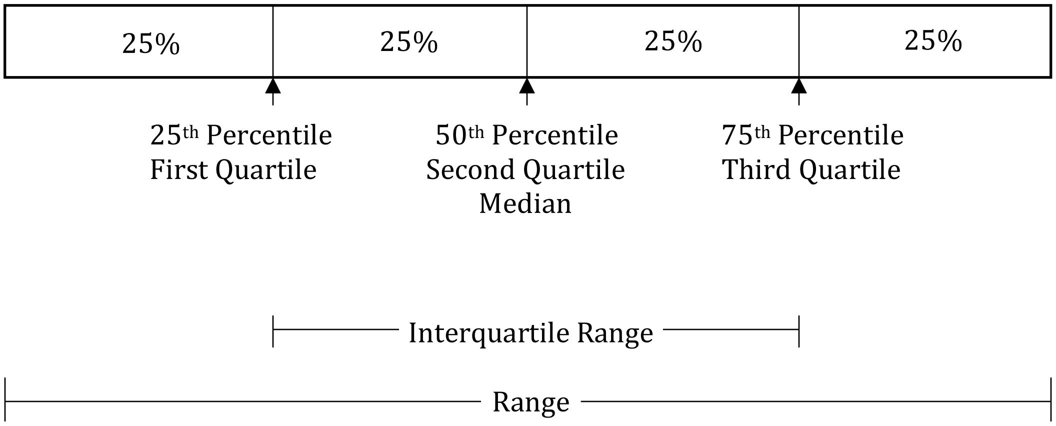
A more informative way to visualize the variability of a distribution is the **boxplot** (also called a **box and whiskers plot**). The boxplot utilizes the median and the range, and also includes a central box delineating the 25th and 75th percentiles.

<u>*Boxplot*</u> *– A summary of a distribution which includes the median, a central box with the 25th and 75th percentiles as limits, and the range. Another name for a boxplot is a box and whiskers plot.*

<u>*Box and whiskers plot*</u> *– Another name for a boxplot.*

Recall that the median divides a distribution so that half (50%) of the scores are below it, and half (50%) are above it. It is, therefore, at the 50th percentile. We are now going to divide a distribution into four regions so that each consists of a quarter (25%) of the scores. If the distribution is rectangular, the result would appear as is shown in top portion of Figure 3.8.

Figure 3.8 Illustration of the Relationship of the Range, Interquartile Range, Median, Percentiles and Quartiles for a Rectangular Distribution



As an example of a boxplot, assume that ten students took an exam. Their scores are shown in Table 3.10.

Table 3.10 Ten Scores on an Exam

68	70	72	73	74	76	79	83	94	97

The range of these scores would be found by subtracting the value of the lowest score from the highest, which would be 97 – 68 which is equal to 29. The median is found by determining the value corresponding to the (N + 1) / 2 position. This would be (10 + 1) / 2 which equals 5.5. There is no 5.5th value, so we find the mean of the 5th and 6th values, which would be (74 + 76) / 2 which equals 75. And we know that the median is at the 50th percentile. The value of the score at the 50th percentile is also called the **second quartile** (Figure 3.8).

Second quartile – The value of the score at the 50th percentile in a distribution. It is the median.

The value with 25% of the distribution below it (25th percentile) would correspond to the median of the bottom half of the distribution. In other words, for our example it is the median of the scores of 68, 70, 72, 73 and 74. The median of these five scores is 72. This value is at the 25th percentile, and is called the **first quartile** (Figures 3.8 and 3.9).

First quartile – The value of the score at the 25th percentile in a distribution.

The value with 75% of the distribution below it (75th percentile) would correspond to the median of the upper half of the distribution. In other words, for our example it is the median of the scores of 76, 79, 83, 94 and 97. The median of these five scores is 83. This value is at the 75th percentile, and is called the **third quartile** (Figures 3.8 and 3.9).

Third quartile – The value of the score at the 75th percentile in a distribution.

In other words, the central 50% of the exam scores would fall between 72 (25th percentile or first quartile) and 83 (75th percentile or third quartile). This central 50% of the distribution is called the **interquartile range** or IQR (Figures 3.8 and 3.9). It is also represented by the 'box' in Figure 3.9. Lines, called **whiskers**, extend from the edges of this box (the 25th and 75th percentiles) to the limits of the data (the range) (Figure 3.9). A boxplot is also often drawn vertically, which is shown in Figure 3.10. (Note that some descriptions of the boxplot emphasize that the whiskers do not include any data points that are identified as outliers in the distribution. As this is an introductory text, a discussion of outliers has not been included.)

Interquartile range (IQR) – A measure of variability based upon the median that includes the middle 50% of the data. It is the range of values in a distribution between the 25th and 75th percentiles.

Whisker – In a boxplot, a line extending from an edge of the box (either the 25th or 75th *percentiles) to the limits of the data. The two whiskers thus extend as far as the range of the data.*

Figure 3.9 Boxplot of the Student Exam Scores

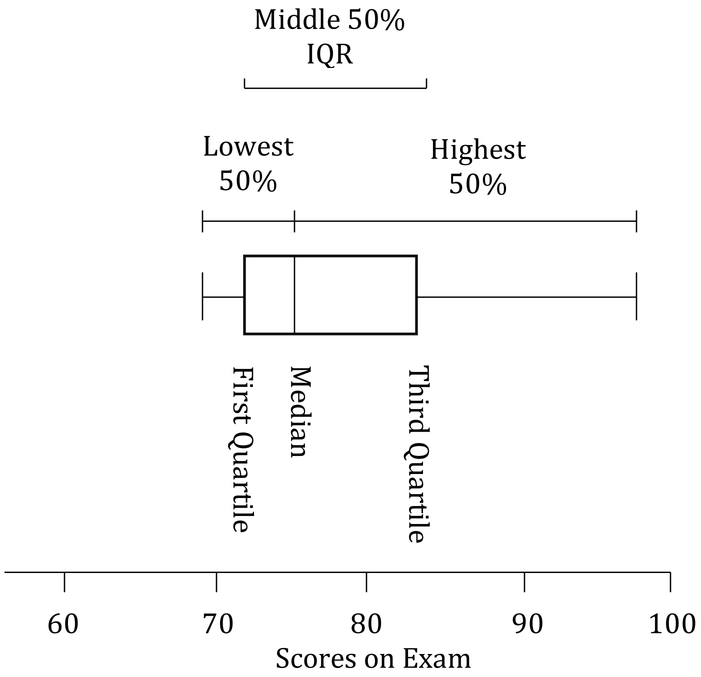
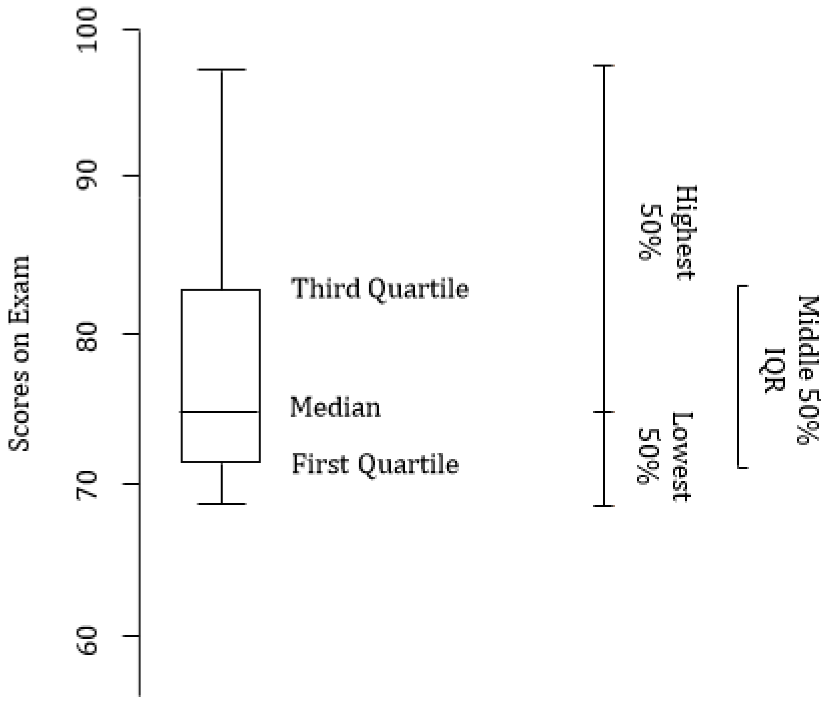


Figure 3.10 Alternative Presentation of the Boxplot of the Student Exam Scores



A great deal of information is conveyed by Figure 3.9 or 3.10. First, while the range of the scores extends from 68 to 97, half of the scores occur within a much smaller 'box' with limits of 72 and 83. Second, the median, which is the value with half of the incomes above it and half below it, is not located near the middle of the distribution of exam scores. Instead, with a value of 75 it is considerably below the physical middle of the range of values. This indicates that the distribution is positively skewed, with the larger tail for higher scores. This conclusion is confirmed by the third observation which is that the distance from the median to the upper limit of the IQR is greater than the distance from the median to the lower limit of the IQR. If the distribution had been symmetrical, the median would have been located at the physical middle of the IQR.

In the literature you will often see figures with more than one boxplot. This allows a quick, informative comparison of multiple sets of data.

It is also important to note that an advantage of the interquartile range as a measure of variability compared to the range is that, unlike the range, the interquartile range is not sensitive to a change in an extreme score. It is thus more stable, which is a beneficial characteristic for a statistical measure.

Once the first quartile and third quartile have been determined it is easy to calculate another commonly used measure of variability, the **semi-interquartile range**. The semi-interquartile range, or SIQR, is simply half of the interquartile range. In other words,

$$\text{SIQR} = \frac{\textbf{interquartile range}}{\textbf{2}}$$

$$= \frac{\textbf{75th percentile} - \textbf{25th percentile}}{\textbf{2}}$$

In our example:

$$\text{SIQR} = \frac{\mathbf{83-72}}{\mathbf{2}}$$

$$= \frac{\mathbf{11}}{\mathbf{2}}$$

$$= 5.5$$

The semi-interquartile range is an informative and commonly used measure of variability for interval or ratio data, particularly when the distribution is skewed.

Semi-interquartile range (SIQR) – A commonly used measure of variability, particularly for skewed data. It is equal to half of the interquartile range.

To be certain you understand the calculations that go into creating a boxplot we will now assume that one additional student's score is included in the previously described exam. This student scored 67. There would now be eleven scores on the exam (Table 3.11).

Table 3.11 Eleven Scores on an Exam

67	68	70	72	73	74	76	79	83	94	97

The range of these scores would be found by subtracting the lowest score from the highest, which would now be 97 – 67 which is equal to 30. The median is found by determining the value corresponding to the (N + 1) / 2 position. This would be the value corresponding to (11 + 1) / 2, which equals the 6th position. This equals a score of 74. This score is at the 50th percentile, which is also called the second quartile.

The value with 25% of the distribution below it (25th percentile or first quartile) would correspond to the median of the bottom half of the distribution. When there is an uneven number of data points in the total distribution, as is the case in our example, the overall median is not included when calculating the first or third quartiles. In other words, for the current example the first quartile would be found by calculating the median of the scores of 67, 68, 70, 72 and 73. The median of these five scores is 70.

The value with 75% of the distribution below it (75th percentile) would correspond to the median of the upper half of the distribution. As was just noted, when there is an uneven number of data points in the total distribution, the overall median is not included in the calculation of the third quartile. In other words, for our example the third quartile would be the median of the scores of 76, 79, 83, 94 and 97. The median of these five scores is 83.

Thus, the central 50% of the eleven exam scores would fall between 70 (25th percentile or first quartile) and 83 (75th percentile or third quartile). This central 50% of the distribution would be the interquartile range or IQR.

With this information we could now construct a boxplot. Hopefully you agree that a boxplot is a useful way to describe a set of data, and that the calculations needed in order to create a boxplot are not difficult.

Finally, the semi-interquartile range, or SIQR, which is simply half of the interquartile range, would be:

$$\begin{aligned}\text{SIQR} &= \frac{\textbf{interquartile range}}{\mathbf{2}}\\ &= \frac{\textbf{75th percentile} - \textbf{25th percentile}}{\mathbf{2}}\end{aligned}$$

In our example:

$$\begin{aligned}\text{SIQR} &= \frac{\mathbf{83-70}}{\mathbf{2}}\\ &= \frac{\mathbf{13}}{\mathbf{2}}\\ &= 6.5\end{aligned}$$

When The Data Are Normally Distributed

Much of this text deals with interval or ratio data that are normally distributed. A boxplot can also be a very informative way to present these data. However, much of the remainder of the book is based upon a critical concept, the deviation. In everyday use, 'deviant' indicates that there is a difference, and the word has a rather negative connotation. In statistics, it just indicates a difference, usually from the mean. There isn't any value judgment. In fact, we are all deviant. No one has the mean score for all traits. Each of us is a little heavier, or shorter, or smarter, or quieter, or happier than the mean.

A related concept, the standard deviation, can be thought of as *the amount that a score is expected to vary from its mean*. However, before proceeding with a discussion of deviations, we need to make a short detour to understand the difference between a population and a sample.

Population And Sample: Statisticians' Way Of Saying 'All' And 'Some'

In statistics, the entire group that is of interest is called a **population**. Any part or sub–set of a population is called a **sample**. An example should make the distinction clear. Assume that you are a teacher. If the entire group that is of interest to you consists of the members of your class, then the members of your class would be a population and any sub-set of it, such as the students who are sitting in the front row, would be a sample. However, if all of the students who attend your school are the group that is of interest, then your class would now be a sample of this much larger group. In other words, whether a group should be considered a population or a sample depends upon the specific situation.

Population – The entire group that is of interest.

Sample – A subset of a population.

Introducing The Variance And Standard Deviation Of A Population

Now we can return to our discussion of the standard deviation. This will at first seem strange, but the best way to introduce the standard deviation is to begin with a discussion of a closely related statistical measure, the **variance**. Both the standard deviation and the variance are measures of variability. The variance is defined as the average of the sum of the squared deviations from the mean. This is, admittedly, not a particularly enlightening definition. Fortunately, this is a case where the mathematical equation is much clearer than the verbal definition, even though the symbols may appear peculiar at first. For a population (not a sample), the equation for the variance is:

Variance of a population $= \sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$

Hopefully you recognize that this is actually a very simple equation. It is just necessary to break it down into its parts and learn the meaning of the new symbols. First, the symbol σ^2 (pronounced sigma squared) is just another way of saying that we are dealing with a variance of a population. (Note that σ is the lower case of the Greek letter sigma. This needs to be distinguished from Σ, the upper case of sigma, which is the symbol for summation. And now you are probably beginning to gain an appreciation of why the phrase 'it is all Greek to me' is sometimes linked with statistics.) Next, the $(X - \mu)$ section of the equation indicates that we are to take a score, symbolized by the letter X, and subtract the population mean, symbolized by the Greek letter μ (pronounced mu). This difference between a score and its mean is called a **deviation**. The $(X - \mu)^2$ part of the equation indicates that we are to square this deviation. Next, the numerator, $\Sigma(X - \mu)^2$, indicates that we are to sum the squared deviations of all of our scores. This term is called, appropriately, the 'sum of the squared deviations'. Finally, the / N indicates that we are to divide the sum of the squared deviations that we just found by the total number of scores. This is a complex paragraph, but I am confident that an example will make the steps clear.

Variance – A measure of variability; the average of the sum of the squared deviations of scores from their mean. The symbol for the population variance is σ^2.

Deviation – The difference between a score and its mean. Thus, with population data the deviation equals $X - \mu$.

Let us assume that you are a member of a very select group that consists of only three individuals. (I have chosen an unrealistically small data set to assist you in understanding the calculations.) Someone is interested in determining how the group is doing and collects data from each person on a quiz. The scores are listed in Table 3.12. As this is the only group that is of interest these three individuals constitute a population, not a sample.

We can now proceed to calculate the variance. The first step is to calculate the population mean. The sum of the three scores is 24. To find the mean we divide the sum of the scores by the total number of scores, which in this case is three. The mean thus equals 24 / 3, which is 8. We now turn to finding the deviations. The deviation of the first quiz score, symbolized by $(X - \mu)$, would equal 6 – 8. This is –2, and it is indicated in the first entry of the third column by being bolded. When we square –2, we obtain 4, the deviation squared, which is the first entry in the fourth column. We would then proceed to the second and third quiz scores.

Table 3.12 Initial Steps in Calculating the Variance

Subject	Score	Deviation $(X - \mu)$	Deviation Squared $(X - \mu)^2$
1	6	**–2**	4
2	8	0	0
3	10	2	4
Σ =	24	0	8

Calculating the variance is now just a matter of substituting into the equation:

$$\text{Variance of a population} = \sigma^2 = \frac{\Sigma\,(X - \mu)^2}{N}$$

$$= \frac{8}{3}$$

$$= 2.67$$

The good news is that we have just calculated a variance. Unlike the range, which is based solely on the two most extreme scores and is thus unstable, the variance is affected by all of the scores and is, therefore, more stable. This is a good feature. The bad news is that the variance is not a particularly useful descriptive statistic. The reason is that the variance is measured in squared units, as is indicated by the symbol σ^2. In other words, while the variance is providing a measure of variability, in this case it is 2.67 points squared, which is probably not the easiest concept to grasp.

There are two obvious solutions to the problem of the variance being measured in squared units. First, you might suggest that we simplify the entire process and base our measure of variability on the deviation scores. In other words, if we do not square the deviations, then our measure of variability will not be in squared units. Unfortunately, while it is true that you will have solved one problem, you will have created another. If you refer to Table 3.12, you will see that the sum of the column of deviations equals zero. This will always be the case. No matter what set of numbers is being examined, the sum of the deviations from the mean will always equal zero. Clearly, the sum of the deviations from the mean will not work as a measure of variability.

The other obvious solution to our problem with the squared units of the variance is to simply take the square root. This puts the measure of variability back into the original units. The result is a measure of variability known as the **standard deviation**, which has the symbol σ. In other words:

Standard deviation = √ variance

and

Variance = (standard deviation)^2

Referring to our example of three scores on a quiz (Table 3.12), the standard deviation would equal the square root of 2.67 points squared, which is 1.63 points. This measure of variability is back in the original units of measurement, points on the quiz. And you will see that it is a very useful measure for it indicates how much we expect a score to vary from its mean. We can, therefore, succinctly summarize the central tendency and variability of a set of interval or ratio scores by providing the mean and the standard deviation. With our example of the three quiz scores, the mean is 8 points and the standard deviation is 1.63 points.

Standard deviation – A measure of variability; the expected deviation of a score from its mean. It is defined as the square root of the variance. The symbol for the population standard deviation is σ.

You will see in this text that with interval and ratio data the standard deviation is the most frequently used measure of variability with descriptive statistics, while the variance is the most frequently used measure of variability with inferential statistics.

A Few More Symbols

We have just learned that the difference between a score and its population mean (X – μ) is called a deviation. The concept of a deviation is used so commonly in statistics that it is given its own symbol, *x*. (Note that a capital X is used to represent a score and a lower case *x* is used to represent a deviation. It is important to keep this distinction clear.) Similarly, the **sum of the**

squared deviations [$\Sigma(X - \mu)^2$] is used so commonly that it also has its own abbreviation, SS. It should be clear that another way to express SS is Σx^2.

<u>Sum of the squared deviations</u> – For a population, it is equal to $\Sigma(X - \mu)^2$ or Σx^2. It is often abbreviated as 'sum of squares' which is shortened even further to SS.

Since the sum of the squared deviations can be abbreviated in a variety of ways, it follows that the equation for the variance can also be written in a number of forms. We have already seen that the variance of a population has the symbol σ^2, and is defined as $\Sigma(X - \mu)^2$ / N. In addition, it was just pointed out that the sum of the squared deviations [$\Sigma(X - \mu)^2$] is abbreviated as SS. Therefore, the population variance also could be written as SS / N. Further, as $\Sigma(X - \mu)^2$ is also abbreviated as Σx^2, the population variance can be written as Σx^2 / N. And the standard deviation is equal to the square root of each of these forms of the variance equation, as is indicated in Table 3.13.

Table 3.13 Equations for the Population Variance and Standard Deviation

<u>Variance</u>	<u>Standard Deviation</u>
$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$	$\sigma = \sqrt{}\frac{\Sigma(X - \mu)^2}{N}$
$\sigma^2 = \frac{SS}{N}$	$\sigma = \sqrt{}\frac{SS}{N}$
$\sigma^2 = \frac{\Sigma x^2}{N}$	$\sigma = \sqrt{}\frac{\Sigma x^2}{N}$

A further example may assist in clarifying the use of these symbols and equations.

Let us assume that we have an interest in the heights of basketball players on a college team. Specifically, we want to determine the mean and standard deviation of the heights of the five starting players. Their heights in inches are listed in Table 3.14. As these are all of the players that we are interested in, this group of five individuals is a population.

Table 3.14 Heights of Five Basketball Players in Inches

Player	Height (X)
1	70
2	72
3	76
4	80
5	<u>81</u>
	$\Sigma X = 379$

The mean of the five heights would be found using the following equation:

$$\text{Mean} = \frac{\Sigma X}{N}$$

$$\mu = \frac{379}{5}$$

$$\mu = 75.8 \text{ inches}$$

We can now use the equation, $\sigma = \sqrt{[(\Sigma(X - \mu)^2) / N]}$, from Table 3.13, to find the standard deviation. The first step is to find the deviation, $(X - \mu)$, which can also be written as *x*, for each of the five heights. We then square these values. This outcome can be written as $(X - \mu)^2$ or x^2. These steps are illustrated in Table 3.15.

Table 3.15 Initial Steps in Calculation of the Standard Deviation

	X	$(X - \mu)$ or x	$(X - \mu)^2$ or x^2
	70	(70 – 75.8) = –5.8	33.64
	72	(72 – 75.8) = –3.8	14.44
	76	(76 – 75.8) = 0.2	0.04
	80	(80 – 75.8) = 4.2	17.64
	81	(81 – 75.8) = 5.2	27.04
Σ	= 379	= 0	= 92.80

As a check on our arithmetic, we confirm that the sum of the deviations, $\Sigma(X - \mu)$, which can also be written as Σx, is zero. We now proceed to find the sum of the squared deviations, $\Sigma(X - \mu)^2$, which can also be written as Σx^2 or SS. This is equal to 92.80 inches squared (Table 3.15).

The next step, as is evident from the equations in Table 3.13, is to divide 92.80 inches squared by N. As a result, we find that $\Sigma(X - \mu)^2 / N$ (which is equivalent to SS / N or $\Sigma x^2 / N$) is equal to 92.80 divided by 5, which in turn equals 18.56 inches squared. We have just found the variance of the heights of the basketball players. Notice again that this variance is measured in inches squared. This is not a particularly meaningful number, so we now take the square root, which will give us the standard deviation of 4.31 inches.

We have just used the three equations for the variance and standard deviation in Table 3.13. They are simply different ways to write the definitional equations for these two measures of variability.

Reporting The Calculated Values Of The Mean And Standard Deviation

If we wanted to report the results, we would say, "The mean of the heights of the five basketball players, in inches, as well as the standard deviation were calculated ($\mu = 75.8$, SD = 4.31)." This would indicate to the reader that the players were tall, 75.8 inches or almost 6 feet 4 inches on average, and that the typical or 'standard' difference between each height and the mean was slightly over 4 inches. A great deal of information has been conveyed with only two numbers. This efficient summary of data is the goal of descriptive statistics.

Progress Check

Assume a population consisting of four soccer players scores 6, 8, 9 and 12 goals during a season.

1. What is the mean?
2. What is the variance?
3. What is the standard deviation?

Answers: 1. 8.75 goals 2. 4.69 goals squared 3. 2.17 goals

Effect On The Variance And Standard Deviation Of Adding Or Multiplying Every Score In A Distribution By A Constant

There are times when a constant number is added to every score in a set of data, such as when a professor curves the scores on a test. It is important to understand how adding or multiplying by a constant will affect the mean and standard deviation of the set of scores. Using our previous example, imagine that each basketball player started playing on stilts that were 12 inches high; then each player's height would increase by 12 inches. This would, in turn, increase the mean height by 12 inches. But how would it affect the standard deviation? The situation is summarized in Table 3.16.

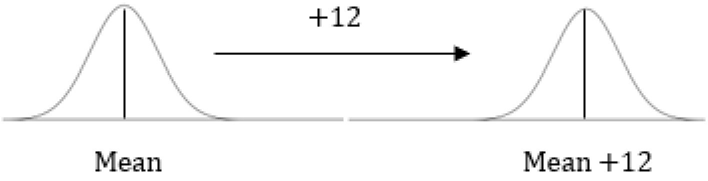
Table 3.16 Illustration of the Effect of Adding a Constant to Every Score

	Original Score	New Score	Deviation	Deviation Squared
	OS	OS + 12 = X	$(X - \mu)$ or x	$(X - \mu)^2$ or x^2
	70	82	(82 – 87.8) = –5.8	33.64
	72	84	(84 – 87.8) = –3.8	14.44
	76	88	(88 – 87.8) = 0.2	0.04
	80	92	(92 – 87.8) = 4.2	17.64
	81	93	(93 – 87.8) = 5.2	27.04
Σ	= 379	= 439	= 0	= 92.80
μ	= 75.8	= 87.8		

The original mean height was 379 / 5, or 75.8 inches. The mean of the new heights is 439 /5 or 87.8 inches, 12 inches greater than the original mean. We then confirm, using the new mean, that $\Sigma(X - \mu)$ equals zero. Next, we note that the sum of $(X - \mu)^2$, which can also be written as Σ(X –

$\mu)^2$, Σx^2 or SS has not changed. It is still equal to 92.80 inches squared. Since the N, which is 5, has not changed, this will lead to a variance and standard deviation that are also the same as was calculated with the previous example. In other words, *if you add a constant value to every score in a set of data, the mean will increase by this constant but the standard deviation and variance will not change.* All that you have done by adding a constant value to every score is to shift the distribution to the right on the number line, as is indicated in Figure 3.11. The mean of the distribution increases by the constant, but as the shape of the distribution and the spread of the scores do not change, neither do the variance or the standard deviation.

Figure 3.11 Effect of Adding a Constant to Each Score



Similarly, if you subtract a constant value from every score in a set of data the mean will decrease by the amount of the constant but, once again, the shape of distribution as well as the variance and standard deviation will not be altered. You can verify that the variance and standard deviation do not change by using the data for basketball players' heights and subtracting a constant. But what happens if you multiply or divide each score by a constant?

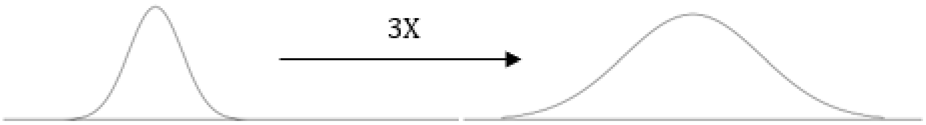
The situation that would result from multiplying each basketball player's height by 3 is indicated in Table 3.17 (This would result in a very tall team!).

Table 3.17 Illustration of the Effect of Multiplying Each Score by a Constant

	Original Score OS	New Score 3(OS) = X	Deviation $(X - \mu)$ or x	Deviation Squared $(X - \mu)^2$ or x^2
	70	210	(210 – 227.4) = –17.4	302.76
	72	216	(216 – 227.4) = –11.4	129.96
	76	228	(228 – 227.4) = 0.6	0.36
	80	240	(240 – 227.4) = 12.6	158.76
	81	243	(243 – 227.4) = 15.6	243.36
Σ	= 379	= 1137	= 0	= 835.20
μ	= 75.8	= 227.4		

The new mean of the heights would be found by dividing the total of the heights, 1137 inches, by 5. This would equal 227.4 inches, three times the original mean which was 75.8 inches. In other words, multiplying each player's height by three also results in a mean height that is three times as large as the original mean height. To find the effect on variability of multiplying each height by three we need to find the sum of the squared deviations which, you recall, can be written as $\Sigma(X - \mu)^2$, Σx^2 or SS. Before doing so, we check that $\Sigma(X - \mu)$, which can also be written as Σx, is zero. Having determined this, we proceed to find that $\Sigma(X - \mu)^2$ equals 835.20 inches squared. Substituting into the equation $\sigma^2 = \Sigma(X - \mu)^2 / N$ from Table 3.13, we obtain our variance which equals 835.20 / 5, or 167.04 inches squared. The standard deviation is the square root of the variance, which in this case would be $\sqrt{}167.04$, or 12.92 inches. This is, except for a small rounding difference, three times the standard deviation of 4.31 inches that we obtained previously. In other words, *if all of the scores in a set of data are multiplied by a constant, the mean and the standard deviation (but not the variance) will also be multiplied by that constant.* This can be illustrated in Figure 3.12, which shows that our distribution not only moved to the right due to the value of each score tripling, it also became three times as spread out. You are encouraged to divide each score in a set of data by a constant and verify that in this case the mean and the standard deviation (but not the variance) will each be divided by your constant.

Figure 3.12 Effect of Multiplying Each Score by Three



Comparing Measures Of Variability For Populations and Samples

Thus far in this chapter we have dealt with the variance and standard deviation (SD) of a population. Fortunately, the situation is virtually identical if you are dealing with a sample. As you recall, a sample is a subset of a population. In our example with the basketball players that began with Table 3.14 we were only interested in the heights of the five starting players. They thus constituted a population. Let us assume, instead, that there were 20 basketball players on a team and our 5 players were chosen from this group. Our 5 players would now constitute a sample of this population of 20 basketball players. If we remain interested in simply summarizing the data by calculating the mean, variance and standard deviation you will see that very little changes.

Nevertheless, when discussing data it is important to keep the distinction between a population and a sample clear. Measures of characteristics of a population, such as its mean and standard deviation, are called **parameters**. Measures of characteristics of a sample, such as its mean and standard deviation, are called **statistics**. As this book deals with a discipline called statistics, not parameters, it should be obvious that we will be working with samples much more often than populations. In order to assist the readers (and writers) of statistics texts and scientific articles in keeping the distinction between a population and a sample clear, different symbols are used. Table 3.18 lists some of the common symbols. It may be helpful to recognize that population parameters are usually signified by Greek letters while sample statistics are signified by Roman letters.

Parameter – A measure of a characteristic of a population, such as its mean or its variance.

Statistic – A measure of a characteristic of a sample, such as its mean or its variance.

Table 3.18 Symbols Used when Describing Population Parameters and Sample Statistics

	Population Parameter	Sample Statistic
Size of Data Set	N	n
Mean	μ	M
Variance	σ^2	s^2
Standard Deviation	σ	s

We have previously defined the deviation of a score from its *population mean* as $X - \mu$. From Table 3.18, it is evident that the symbol for the mean changes when we are dealing with a sample. Consequently, the deviation of a score from its *sample mean* would be written as $X - M$. Similarly, from Table 3.13 you will see that the equation for the *population variance* is $\sigma^2 = \Sigma(X - \mu)^2 / N$. It would be reasonable to assume that by substituting the symbols listed in Table 3.18 we would then have the equations for the sample variance and the sample standard deviation.

It is important to note that if this were the case then with descriptive statistics while the symbols used in the equations for populations and samples would have changed, they would lead to identical outcomes. This would be logical because with descriptive statistics we are only interested in the data set we are currently examining. In other words, if we are dealing with the descriptive statistics of a sample, the observed data are all that we are concerned with and thus, conceptually, these data would essentially be treated in the same way as if they constituted a population. In other words, it would be reasonable to assume that the standard deviation *describing* a set of data would be the same regardless of whether we are dealing with a population or a sample. However, later in the text you will see that with inferential statistics it is necessary to make a minor change to these equations. And inferential statistical procedures are used much more commonly than descriptive procedures. Thus, it is not surprising that we would want to avoid the use of two sets of almost identical equations which lead to very similar, but nevertheless somewhat different, results. Consequently, when calculating the variance or standard deviation of a sample we commonly use the equations developed for inferential statistics even if we are actually asking a descriptive question. Fortunately, the only difference is that we now divide by n – 1 instead of n. (The reason for this modification will be explained later in the text.) Thus the equation for the sample variance becomes $s^2 = (\Sigma(X - M)^2) / (n - 1)$. And the equation for the sample standard deviation would then be $s = \sqrt{}[(\Sigma(X - M)^2) / (n - 1)]$. The different forms of the equations used to calculate the standard deviation, for both populations and samples, are provided in Table 3.19.

Table 3.19 Equations for the Standard Deviation when Describing Populations and Samples

Population	Sample
$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$	$s = \sqrt{\frac{\Sigma(X - M)^2}{n - 1}}$
$\sigma = \sqrt{\frac{SS}{N}}$	$s = \sqrt{\frac{SS}{n - 1}}$
$\sigma = \sqrt{\frac{\Sigma x^2}{N}}$	$s = \sqrt{\frac{\Sigma x^2}{n - 1}}$

So, what is the effect of this change from using N in the denominator to using n – 1? In Table 3.14 we were given the heights of 5 basketball players. As these were the only players of interest, they constituted a population. We subsequently found that the variance was equal to 18.56 inches squared and the standard deviation for these data was equal to 4.31 inches. If our 5 players were instead a sample from a population of basketball players, the mean and the sum of the squared deviations from the mean would not change. However, we would now divide the sum of the squared deviations from the mean, which we calculated to be 92.80 inches squared, by n – 1 instead of N. As a result, we find that the variance, which is now written as $\Sigma(X - M)^2 / (n - 1)$, is equal to 92.80 divided by 5 – 1, which in turn equals 23.20 inches squared. And to find the standard deviation we would take the square root, which will give us a value of 4.82 inches. These values differ substantially from 18.56 inches squared and 4.31 inches which we calculated previously. However, a sample consisting of only five subjects is unusually small, and with larger sample sizes the difference between dividing by n – 1 instead of N quickly becomes negligible.

Conclusion

This chapter has begun the review of the descriptive statistics utilized with interval or ratio data. It was noted that there are a variety of procedures to assist in gaining an overview, including the stem-and-leaf display, the histogram, the frequency polygon and the boxplot. A distinction was also made between skewed and symmetrical distributions. If the data are clearly skewed the appropriate measures for central tendency and variability would be the median and the interquartile or semi-interquartile range, respectably. However, as this text will be emphasizing the analysis of normally distributed data, and these are symmetrically distributed, the descriptive statistics for central tendency and variability that we will most commonly be using will be the mean and the standard deviation. The relation between these statistics and the statistics utilized with nominal and ordinal data are summarized in Table 3.9. Though the calculations involved with finding the descriptive statistics for normally distributed interval or ratio data are somewhat more involved, you will see in the next chapter that the amount of information gained is substantially greater.

Glossary of Terms

Bell-shaped curve – A symmetrical distribution in which the highest frequency scores are located near the middle and the frequency drops the farther a score is from the middle.

Box and whiskers plot – Another name for a boxplot.

Boxplot – A summary of a distribution which includes the median, a central box with the 25th and 75th percentiles as limits, and the range. Another name for a boxplot is a box and whiskers plot.

Deviation – The difference between a score and its mean. Thus, with population data the deviation equals $X - \mu$. The symbol for a deviation is x.

First quartile – The value of the score at the 25th percentile in a distribution.

Frequency polygon – A graphic presentation for use with interval or ratio data. It is similar to a histogram except that the frequency is indicated by the height of a point rather than the height of a bar. The points are connected by straight lines.

Histogram – A graph used with interval/ratio data. As with the bar graph, frequencies are indicated by the length of the associated bars. However, as the data are continuous in a histogram the bars are positioned side-by-side.

Interquartile range (IQR) – A measure of variability based upon the median that includes the middle 50% of the data. It is the range of values in a distribution between the 25th and 75th percentiles.

Leaf – The last digit(s) of a score. With a stem-and-leaf display each leaf is paired with the appropriate stem value and the leaves are listed in ascending order in each row of the display.

Mean – A measure of central tendency for use with interval or ratio data. It is what is commonly called an average. The mean is the sum of the scores divided by the number of scores.

Negatively skewed – A nonsymmetrical distribution in which the tail pointing to the left is larger than the tail pointing to the right.

Normal distribution – A specific, bell-shaped distribution. Many statistical procedures assume that the data are distributed normally.

Parameter – A measure of a characteristic of a population, such as its mean or its variance.

Population – *The entire group that is of interest.*

Positively skewed – *A nonsymmetrical distribution in which the tail pointing to the right is larger than the tail pointing to the left.*

Range – *A measure of variability. It is commonly defined as the value which is obtained when the lowest score is subtracted from the highest score.*

Real limits – *With interval or ratio data, the actual limits used in assigning a measurement. These are halfway between adjacent scores, and are called the upper and lower real limits.*

Sample – *A subset of a population.*

Second quartile – *The value of the score at the 50th percentile in a distribution. It is the median.*

Semi-interquartile range (SIQR) – *A commonly used measure of variability, particularly for skewed data. It is equal to half of the interquartile range.*

Skewed – *A distribution in which one tail is larger than the other. As a result, the distribution is not symmetrical.*

Standard deviation – *A measure of variability; the expected deviation of a score from its mean. It is defined as the square root of the variance. The symbol for the population standard deviation is σ.*

Statistic – *A measure of a characteristic of a sample, such as its mean or its variance.*

Stem – *With a stem-and-leaf display, a list of the different values of the data once the last digit(s) of each score is removed.*

Stem-and-leaf display – *A commonly used summary of interval or ratio data in which each original score is separated into two parts, a stem and a leaf.*

Sum of the squared deviations – *For a population, it is equal to $\Sigma(X - \mu)^2$ or Σx^2. It is often abbreviated as 'sum of squares' which is shortened even further to SS.*

Symmetrical distribution – *A distribution in which the right half is the mirror image of the left half. In such a distribution, there is a high score corresponding to each low score.*

Third quartile – *The value of the score at the 75th percentile in a distribution.*

Variance – *A measure of variability; the average of the sum of the squared deviations of scores from their mean. The symbol for the population variance is σ^2.*

Whisker – *In a boxplot, a line extending from an edge of the box (either the 25th or 75th percentiles) to the limits of the data. The two whiskers thus extend as far as the range of the data.*

Questions – Chapter 3

(Answers are provided in Appendix J.)

1. With a stem-and-leaf display, a row with a stem of 12 and leaves of 0, 2 and 6 would

be equivalent to scores of _____, ______, and _____.

a. 12, 14 and 18

b. 0, 24 and 72

c. 120, 122 and 126

d. 1200, 1222 and 1266

2. A frequency polygon is preferred to a histogram when there are (a) _____.

a. Small number of possible X values

b. Many values of Y for each value of X

c. Large number of possible X values

d. Few values of Y for each value of X

3. If a person reports that their height is 5 feet 8 inches, the 'real limits' were actually _____ and _____.

a. 5 feet 7 ½ inches; 5 feet 8 ½ inches

b. 5 feet 7 inches; 5 feet 9 inches

c. 5 feet 8 inches exactly

4. The most obvious difference between a bar graph and a histogram is that _____.

a. The bars touch in a bar graph but are separated in a histogram

b. The bars touch in a histogram but are separated in a bar graph

c. A bar graph is used for interval or ratio data whereas a histogram is only used with nominal data.

d. A bar graph will always have more bars than a histogram will have.

5. What is the mean of 96, 92, 98 and 90?

a. 93

b. 93.5

c. 94.5

d. 94

6. A serious problem with the mean as a measure of central tendency is that _____.

a. It is too difficult to calculate

b. It cannot be used if the set of numbers is large

c. It is affected by extreme scores

7. The two most commonly used measures of variability with normally distributed interval and ratio data are _____ and _____.

a. Standard deviation; variance

b. Range; standard deviation

c. Variance; range

8. In a distribution, the sum of the deviations from the mean will always equal _____.

a. 3

b. 0

c. 6.5

d. It varies depending upon the set of numbers.

9. If you have a distribution consisting of 13 scores, the median would be the _____ score.

a. 1st

b. 3rd

c. 7th

d. 13th

10. What are the median and range for temperatures of 91, 92, 93, and 94?

a. 92; 2
b. 92.5; 3
c. 92.5; 4
d. 93; 5

11. If I am solely interested in the views of my statistics class, I am considering the class to be a _____. However, if I am interested in using the statistics students' views to learn about all college students' opinions, I am considering the class to be a _____.

a. Population; sample
b. Sample; population

12. We use the _____ or _____ to graph interval or ratio data.

a. Histogram; pie chart
b. Bar graph; frequency polygon
c. Pie chart; bar graph
d. Histogram; frequency polygon

13. What is the median and range of heights, measured in inches, of 72, 81, 85, and 91?

a. 83; 19
b. 83; 20
c. 81; 19
d. 85; 20

For questions 14 – 17, assume there are 12 scores:

2 3 3 5 6 7 9 10 12 16 22 60

14. What is the range?

a. 57
b. 58
c. 59
d. 60

15. What is the median (second quartile)?

a. 6
b. 7
c. 8
d. 9

16. What is the value of the first quartile?

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

17. What is the value of the third quartile?

a. 12
b. 13
c. 14
d. 15

For questions 18 – 21, assume we add a score of one to the previous set of numbers. There are now 13 scores:

1 2 3 3 5 6 7 9 10 12 16 22 60

18. What is the range?
a. 57
b. 58
c. 59
d. 60

19. What is the median (second quartile)?
a. 6
b. 7
c. 8
d. 9

20. What is the value of the first quartile?
a. 2
b. 3
c. 4
d. 5

21. What is the value of the third quartile?
a. 12
b. 13
c. 14
d. 15

22. If a distribution is symmetrical, the median will be located _____.
a. closer to the high end of the distribution
b. closer to the low end of the distribution
c. below the mode
d. at the middle of the interquartile range

23. What percentage of scores fall within the interquartile range?
a. 25
b. 50
c. 75
d. 100

24. If the interquartile range is equal to 10, the semi-interquartile range would equal _____.
a. 0
b. 10
c. 20
d. 5

25. The concept of 'real limits' occurs with _____ and _____ measurement scales.
a. Nominal; ordinal
b. Interval; ratio
c. Nominal; interval
d. Ordinal; ratio

26. A bell-shaped curve is _____ and _____.

a. Bimodal; symmetrical
b. Unimodal; skewed
c. Bimodal; skewed
d. Unimodal; symmetrical

27. The mean is the most common measure of central tendency for _____ and _____ measurement scales.
a. Interval; ratio
b. Nominal; interval
c. Ordinal; ratio
d. Nominal; ordinal

28. The difference between a score and its mean is called a _____.
a. Range
b. Real limit
c. Deviation
d. Modality

29. A distribution that is non–symmetrical and has a prominent tail that points to the left is called _____.
a. Negatively skewed
b. Positively skewed
c. Bimodal

30. The variance for the population consisting of the scores 2, 4, 6, 3, and 5 is _____ and the standard deviation is _____.
a. 2.5; 1.58
b. 1.4; 2
c. 20; 4.5
d. 2; 1.4

31. The variance for the sample consisting of the scores 2, 4, 6, 3, and 5 is _____ and the standard deviation is _____.
a. 2.5; 1.58
b. 1.4; 2
c. 20; 4.5
d. 2; 1.4

32. Adding or subtracting a constant (such as 5) to every score in a distribution will change the _____ but not the _____ or _____.
a. Mean; mode; median
b. Mean; variance; standard deviation
c. Standard deviation; variance; mean
d. Variance; mean; standard deviation

33. If all of the scores in a distribution are multiplied by 10, the mean will be _____ times larger and the standard deviation will be _____ times larger.
a. 5; 10
b. 10; 5
c. 10; 10
d. 5; 5

34. The standard deviation will equal 0 when _____.
a. the range is less than 20

b. every score in the distribution is the same
c. the mean is negative
d. the variance is greater than 6

35. If the mean of a distribution is to the right of the median, the distribution is probably ____.
a. Negatively skewed
b. Positively skewed
c. Symmetrical
d. Any of the above are equally likely

36. The more varied the scores in a distribution, ____.
a. The larger the standard deviation will be
b. The smaller the standard deviation will be
c. Variation of scores does not affect the standard deviation

37. For a football team, if the mean yards gained per play were the same for their running and passing plays, but the standard deviation was greater for the passing plays, then _____.
a. They would have a greater chance of making a large gain with a running play
b. They would have a greater chance of making a large gain with a passing play
c. The chance of making a large gain would be the same for a running or a passing play.

38. The variance is equal to the _____.
a. Square root of the standard deviation
b. Standard deviation
c. Square of the standard deviation
d. None of the above

Chapter 4
Describing Interval and Ratio Data – II:
Further Descriptive Statistics Used with Interval and Ratio Data

"The most important questions of life are, for the most part, really only problems of probability."

Pierre Simon, Marquis de Laplace

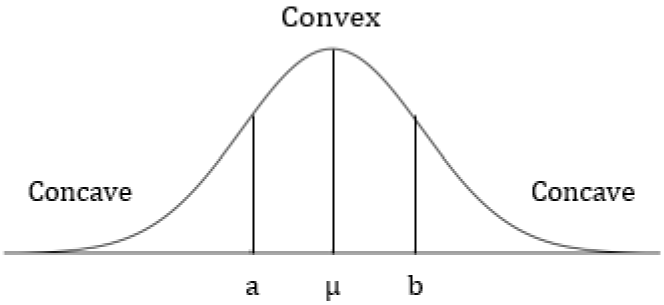
Introduction

In Chapter 3 we learned that the standard deviation is the most commonly used descriptive measure of variability for interval or ratio data that are normally distributed. We have also seen how to find the value of the standard deviation for both populations and samples. In addition, it was noted that unlike the range, the standard deviation makes use of all of the data and will, therefore, tend to be more stable. There are other characteristics of the standard deviation that make it particularly useful as a descriptive statistic.

We previously noted that if interval or ratio data are normally distributed they form a symmetrical, bell-shaped distribution, as is shown in Figure 4.1. If you start at the far left on the graph and follow it to the right, you will see that the direction of the curve changes at point 'a'. To the left of point 'a' the curve is concave, like the inside of a circle; to the right of point 'a' the curve is convex, like the outside of a circle. As you continue to the right from point 'a' the line continues to form a convex curve until you get to point 'b'. At point 'b' the direction changes again and the line begins to form another concave curve. Points 'a' and 'b' are called **inflection points**.

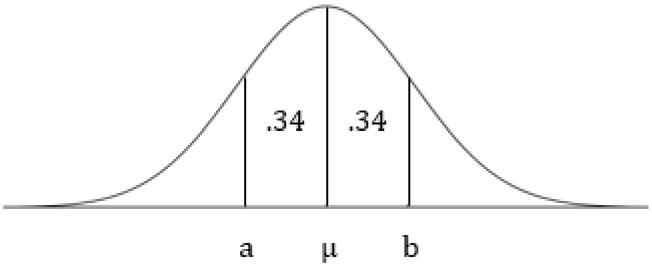
<u>*Inflection point*</u> *– A point on a graph where the curvature changes from concave to convex or from convex to concave.*

Figure 4.1 Inflection Points on a Normal Curve



It should be evident from examining Figure 4.1 that points ‘a’ and ‘b’ are equidistant from the mean. It is also the case that with a normal distribution point ‘a’ is located 1 standard deviation (SD) below the mean and point ‘b’ is located 1 SD above the mean. Further, it has been found that the proportion of the normal curve between point ‘a’ and the mean is approximately 0.34 or about 34%. Similarly, since the curve is symmetrical, the proportion of the curve between point ‘b’ and the mean is also approximately 0.34 or about 34%. Put another way, if there is a normal distribution of 100 scores, then approximately 34 will be in the region between the mean and 1 SD below the mean (point ‘a’) and another 34 scores will be in the region between the mean and one standard deviation above the mean (point ‘b’). In other words, approximately 68% of the total cases will fall within +/–1 SD of the mean when we are dealing with a normal distribution. This relationship between proportions or areas and the normal distribution is illustrated in Figure 4.2.

Figure 4.2 Proportion of the Curve Between the Mean and the Inflection Points

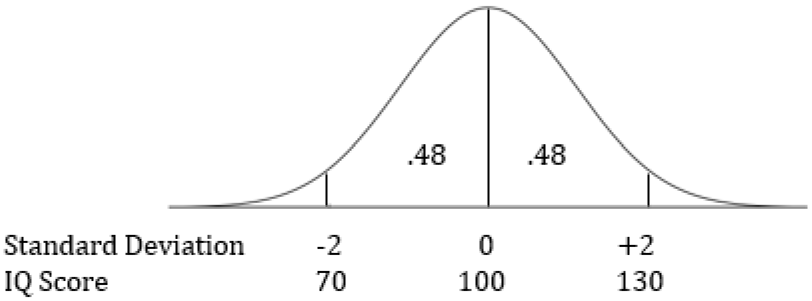


What You Always Wanted To Know About The IQ, But No One Told You

The critical concept to recognize is that so long as the variable is normally distributed there is a precise relationship between the distance (number of standard deviations) a score is from the mean, and the corresponding proportion. For instance, the IQ test is approximately normally distributed and has a mean of 100 and a standard deviation of 15. You now know that if 100 people took the test, then we expect that approximately 34 will score between 85 (1 SD below the mean) and 100 (the mean). Another 34 will score between 100 (the mean) and 115 (1 SD above the mean). Thus, approximately 68, or about two-thirds, of the individuals will have IQ scores between 85 (1 SD below the mean) and 115 (1 SD above the mean).

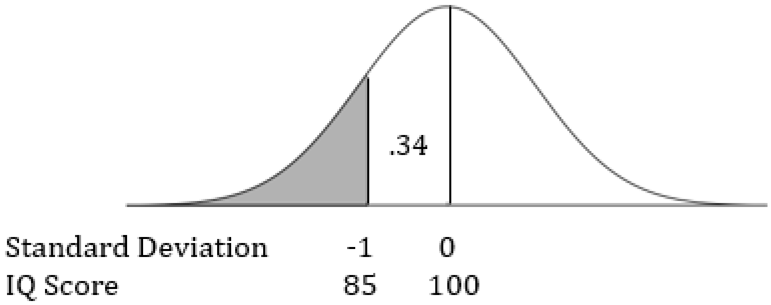
Through the use of calculus we also know the proportion of the distribution between the mean and either plus or minus 2 SD. This proportion is approximately .48, which is shown in Figure 4.3. Using our IQ example, approximately 48 of the 100 individuals who took the test would be expected to fall between 70 (2 SD below the mean) and 100 (the mean). Similarly, approximately 48 of the 100 individuals who took the test would be expected to fall between 100 (the mean) and 130 (2 SD above the mean). In other words, about 96 of the 100 people who took the IQ test would be expected to have scores between 70 and 130. Converting to percentages, slightly over 95% of the cases will fall within 2 SD of the mean. What this indicates is that for *any* normal distribution there is less than a 5% chance that a score will be more than 2 SD from the mean. Later in this book you will learn that the areas associated with all the standard deviations, not just for 1 and 2 SD, have been calculated. For now, we will continue to deal with the values that we have already described.

Figure 4.3 Proportion of the Curve Between the Mean and Plus or Minus Two Standard Deviations



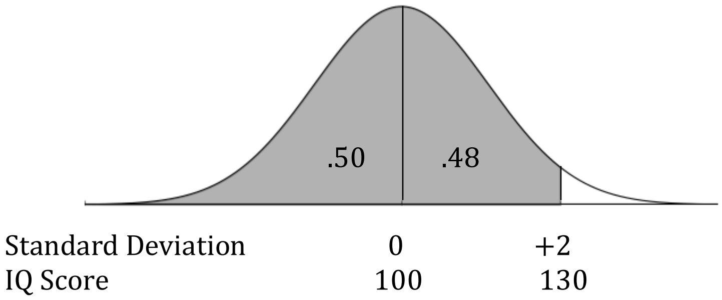
By drawing a new figure and referring to Figures 4.2 and 4.3 it is easy to determine the answers to a number of additional questions. For instance, how many of 100 individuals would be expected to have IQ scores that would fall below 85? In order to determine this number, the first step is to recognize that an IQ of 85 is equivalent to 1 SD below the mean. Next we would draw a figure indicating what is being asked. This is shown in Figure 4.4. By referring to Figure 4.2 we note that 0.34 of the total area falls between the mean and 1 SD below the mean. What we are looking for, however, is the region more than one standard deviation below the mean. Since the normal distribution is symmetrical, the entire area below the mean represents 50%, or 0.50 of the curve. The region that we seek is thus 0.50 – 0.34, which equals 0.16. Since we were asked how many individuals out of 100 would be in this region of the distribution, we multiply 0.16 X 100 and obtain 16. I hope you agree that as long as you draw a figure, working with the proportions and percentages associated with a standard deviation is not particularly difficult.

Figure 4.4 Region of the Curve Below an IQ Score of 85



For our final example, let us find what proportion of the IQ scores would be less than 130. The first step is to recognize that an IQ of 130 is 2 SD above the mean. Then we draw a figure showing the area that is of interest. This is shown in Figure 4.5. From Figure 4.3 we find that the area between the mean and a point 2 SD above the mean is 0.48 of the total area under the curve. Since half, or 0.50, of a symmetrical distribution is below the mean, the region that we are searching for would be equal to 0.48 + 0.50, or 0.98.

Figure 4.5 Region of the Curve Below an IQ of 130



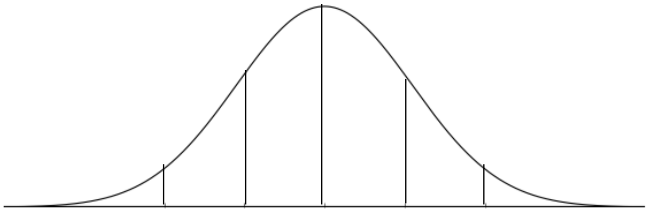
Alternatively, we could have found the proportion of scores above an IQ of 130 and subtracted that amount from the total area under the curve. To do this we could have subtracted 0.48 from the total area to the right of the mean, which is 0.50. This would give us 0.02. We could then subtract this proportion from the total area of the curve, which is 1.0 or 100%. This would give us 1.00 – 0.02 or 0.98, the same value we obtained previously. What this shows is that there may be more than one way to find the desired answer. To be successful, begin by drawing the area you are seeking and, when you finish your calculations, check to make certain that you have provided the answer in the desired form. If the question asks for a proportion, be certain that you answer with the proportion. On the other hand, if the question requests a number of subjects, be sure to convert the proportion or percentage into the desired number.

It should be evident that the standard deviation can be particularly useful when you are dealing with a normal curve. Once the standard deviation is determined, the probabilities associated with different outcomes can be determined. However, the verbal descriptions that we have been using are rather awkward. Expressions such as '2 SD below the mean' or 'the area between the mean and 1 SD above the mean' require some careful attention to be understood. Fortunately, the field of statistics uses a much simpler alternative, the z score.

Later in this chapter you will be given the mathematical definition of a **z score**. For now, just think of it as the number of standard deviations above or below the mean. For example, an IQ score of 115 is 1 SD above the mean. Because it is 1 SD above the mean, it is equivalent to a z score of +1. Similarly, an IQ score of 70 is 2 SD below the mean, which is the same as saying an IQ of 70 has a z score of –2. Thus *the magnitude of the z score is simply the number of standard deviations you are away from the mean and the sign, either positive or negative, indicates the direction.* For instance, a z score of –1 indicates that the point is 1 SD below the mean. In terms of IQ scores, this would be a score of 85. Further, just as the area between the mean and 1 SD above the mean is equal to 0.34 (Figure 4.2), the area between the mean and a z score of +1 is also 0.34. The other areas in Figures 4.2 and 4.3 would also correspond to the associated z scores. Remember, the z score is simply a shorthand way of indicating the number of standard deviations a score is from the mean along with the direction it is from the mean (Figure 4.6).

z score – A conversion of raw data so that the deviation is measured in standard deviation units and the sign, positive or negative, indicates the direction of the deviation.

Figure 4.6 Relationship Between Standard Deviations, IQ Scores and z Scores



Standard Deviation	-2	-1	0	+1	+2
IQ Score	70	85	100	115	130
z Score	-2	-1	0	+1	+2

Progress Check

1. A parameter is a characteristic of a _____ while a statistic is a characteristic of a _____.
2. If there are 100 people, how many would you expect to score between the mean and one standard deviation above the mean?
3. If you score 2.3 standard deviations above the mean, your z score would be _____.

Answers: 1. Population; sample 2. 34 3. +2.3

"It is the nature of probability that improbable things will happen."

Aristotle

Descriptive Statistics – The z Score

We will now be focusing on a further discussion of the descriptive measure known as the z score, which is underlined in Table 4.1.

Table 4.1 Overview of Descriptive Statistics (Summarizing Data)

	Type of Data			
	Nominal (Frequency)	Ordinal (Ranked)	Interval/Ratio (Score)	
Frequency Dist	Bar Graph or Pie Chart	Bar Graph	Histogram or Frequency Polygon	
Central Tendency	Mode	Median	IF NOT NORMAL Median	IF NORMAL Mean (Median – less common)
Variability	– – – –	Range	Interquartile Range	Standard Deviation z Score
Summary Presentation			Stem-and-leaf display and Boxplot	Stem-and-leaf display and Boxplot

A Little History And A Very Impressive Equation

As Table 4.1 shows, the z score is used when there are interval or ratio data. In addition, you have just seen that the z score is particularly useful in those situations in which the data are normally distributed.

The concept of a normal curve was developed when it was noticed that many, but certainly not all, variables tend to exhibit what we call a bell-shaped distribution. For instance, if we were to plot the weights of adult males we would find that many have moderate weights and that progressively fewer would have either extremely low or extremely high weights. This pattern, with a high frequency of events surrounding the mean and progressively fewer occurrences as you move in either direction, attracted the interest of mathematicians. In 1733, Abraham DeMoivre proposed an equation for the normal distribution. It is not something that is likely to be appreciated when first encountered:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/[2\sigma^2]}$$

Fortunately, it is not necessary for you to memorize or even be able to work with this equation. What is important is that you have a modest understanding of the pieces that make up the equation for the normal distribution.

The equation indicates that the likelihood of a score (P(X)) is dependent upon two constants as well as three variables. You may be familiar with the two constants, the natural logarithm 'e', which is approximately 2.18, and π, which is approximately 3.14 (Don't worry, you won't be using these numbers. Remember, we're keeping things simple). The three variables were introduced previously in this text and are the population mean, μ, population standard deviation, σ, and the population variance, σ^2. As there are an infinite number of possible combinations of μ and σ, there are also an infinite number of normal curves. However, they all share a number of characteristics, some of which were introduced in Chapter 3:

1. Unimodal: all normal curves have a single peak or mode.
2. Symmetrical: all normal curves have mirror image shapes to the left and right of the mean.
3. Bell-shaped: all normal curves have shapes that can be described as resembling a bell.
4. The inflection points of this curve occur exactly one standard deviation above and below the mean.
5. All normal curves can be transformed into what can be called a standard normal curve. The standard normal curve has a μ of 0 and a σ of 1.
6. When dealing with the standard normal curve we utilize what are called z scores.

The Standard, Very Important, z Score

The z score was defined as the number of standard deviations that a score differs from its mean. We are now ready for the equation for the z score, which is, $z = (X - \mu) / \sigma$. This equation is quite simple. It defines the value of the z score that corresponds to the value of an individual score (signified with the letter X). Thus any score can be converted into its z score equivalent. From Chapter 3, you will recall that the numerator, $(X - \mu)$, is a deviation. It is the difference between a score and its population mean. The equation for z indicates that this deviation is then divided by the population standard deviation, σ. What this last step accomplishes is to convert a deviation, which was measured in units such as feet or the number of correct answers on an exam, into standard deviation units. You are familiar with doing similar transformations. For instance, if you find that the length of a rug is 15 feet, and then divide by 3 feet, you will have converted the length into yards. Dividing by the population standard deviation, σ, converts the original deviation into standard deviation units. Thus, regardless of what the original measurement unit was, dividing by σ converts the deviation between the original score and the population mean into a deviation measured in standard deviation units. Because of this uniformity of measurement, the z score is also called a **standard score**.

As an example, consider the most commonly used IQ tests, which have a mean of 100 and a standard deviation of 15. Remember, the equation is $z = (X - \mu) / \sigma$. If an individual obtained an IQ score of 145, that person's deviation from the mean would be 145 – 100. The z score would then equal (145 – 100) / 15. This is 45 / 15, or 3 (which is equivalent to +3). Thus, an IQ score of 145 is 3 SD above the mean and is equivalent to a z score of +3. An IQ of 70 would be equal to (70 – 100) / 15. This is –30 / 15, which equals –2. This indicates that an IQ of 70 is 2 SD below the mean and is equivalent to a z score of –2.

It is important to note the positive or negative sign of the z score. Whenever the z score is positive, we are dealing with an original score that is above, or greater than, its mean. Whenever the z score is negative, we are dealing with an original score that is below, or less than, the mean. Thus, an IQ of 145 results in a positive z score because 145 is greater than the mean of the IQ distribution, which is 100. Similarly, we found that with an IQ of 70, the z score is negative because 70 is less than the mean of 100. To reiterate, the magnitude of the z score is simply the number of standard deviations that a score falls from the mean, and its sign indicates the direction.

Standard score – A measure indicating whether a score is above or below the mean as well as how many standard deviations it is from the mean. Also called a z score.

Who Says You Can't Compare Apples And Oranges?

Converting our initial data, which are called **raw scores**, into z scores permits us to make comparisons that would otherwise not be meaningful. For instance, if you scored 9 out of 10 on a music audition and 85 out of 100 on a statistics exam, on which test did you do better? To answer this question your first thought might be to convert one of the scales so it had the same upper limit as the other. For instance, you could multiply your music audition score by 10. This would put it on a 100-point scale. Then, you might conclude that you did better on the music audition since your transformed score, which is now 90, is higher than the statistics score of 85. However, it might also occur to you that this is not a very satisfactory solution. What if most of the scores on the music audition tended to be very high and most of the scores on the statistics exam tended to be lower? In this case it is possible that you had the lowest score on the music audition but the highest score on the statistics exam. Of course, the opposite is also possible. Your score of 90 on the audition might have been the highest score obtained, while your score of 85 on the statistics exam might have been the lowest grade on the exam. As the original measurement scales were different you are now in the situation where you are comparing apples to oranges. What is needed is a standard measurement scale, and this is a situation where the z score can be of great value.

Raw score – Your data as they are originally measured, before any transformation.

If you know the mean and standard deviation of the music audition and statistics exam, and that the two distributions are normal, you can convert each of the raw scores into z scores and then make a meaningful comparison. For instance, if the mean of the music audition scores was 8 and the standard deviation was 0.50, while the mean of the statistics scores was 81 and the standard deviation was 4, then the corresponding z scores could be calculated as shown below:

$$\begin{aligned}\text{z score for music audition} &= \frac{(\mathbf{X} - \mu)}{\sigma}\\ &= \frac{(\mathbf{9 - 8})}{\mathbf{0.50}}\\ &= \frac{\mathbf{1}}{\mathbf{0.50}}\\ &= +2\end{aligned}$$

$$\begin{aligned}\text{z score for statistics exam} &= \frac{(\mathbf{X} - \mu)}{\sigma}\\ &= \frac{(\mathbf{85 - 81})}{\mathbf{4}}\\ &= \frac{\mathbf{4}}{\mathbf{4}}\\ &= +1\end{aligned}$$

These results show that you had a z score of +2 for the music audition and a z score of +1 for the statistics exam. In both cases the z scores are positive, so in each situation you were above the mean. However, you did relatively better on the music audition, for you scored 2 SD above the mean on the music audition and only 1 SD above the mean on the statistics exam.

Clearly, by standardizing the scores, in other words by converting the raw scores into z scores, you are able to make comparisons that would otherwise not be meaningful. Put differently, with z scores you actually can compare apples to oranges!

Just as you can convert a raw score into a z score, it is also possible to do the reverse. For instance, you can find the raw score that is equivalent to a z score of +1.50 on the music audition. This can be accomplished using the original equation for the z score:

$$\text{z score for music audition} = \frac{(\mathbf{X} - \mu)}{\sigma}$$

Substituting the value we are given for the z score, plus the mean and standard deviation from before, leads to the following equation:

$$+1.50 = \frac{(\mathbf{X - 8})}{\mathbf{0.50}}$$

Multiplying each side of the equation by 0.50 gives us:

$$0.75 = \mathrm{X} - 8$$

Adding 8 to each side of the equation leads to the answer:

$$8.75 = \mathrm{X}$$

We conclude that a a z score of +1.50 is equivalent to raw score of 8.75.

If solving this type of equation is awkward for you, the original equation for z can be rearranged so that the X value is presented alone on the left. This is shown below:

$$\mathrm{X} = \mathrm{z}\sigma + \mu$$

Substituting the values for z, μ and σ from the previous example would give us:

$$\begin{aligned}\mathrm{X} &= [(1.50)\ (0.50)] + 8\\ &= 0.75 + 8\\ &= 8.75\end{aligned}$$

This is the same outcome as we obtained before.

To be certain that you feel comfortable converting from z scores to raw scores we will do one more example. What raw score is equivalent to a z score of –3 on the statistics exam? Using the definitional equation for z we would have the following:

$$\text{z score for statistics exam} = \frac{(\mathbf{X} - \mu)}{\sigma}$$

Substituting the values that we know leads to:

$$-3 = \frac{(\mathbf{X - 81})}{\mathbf{4}}$$

We then multiply by 4 to give:

$$-12 = \mathrm{X} - 81$$

To find X, we now add 81 to both sides:

$$69 = \mathrm{X}$$

Alternatively, we could use the version of the equation that has X on the left side of the equation. In this case, we have:

$$\mathrm{X} = \mathrm{z}\sigma + \mu$$

Substituting the values that we know leads to:

$$\begin{aligned} X &= [(-3)\ (4)] + 81 \\ &= -12 + 81 \\ &= 69 \end{aligned}$$

The answer is the same. You can choose whichever equation you find easier to use.

To this point we have converted a raw score into a z score, and we have converted a z score into a raw score. In addition, we have seen that converting raw scores into z scores permits us to make comparisons that would not otherwise be meaningful. We will now see that it can be very useful to convert an entire distribution of raw scores into z scores. Before we do this, it is important to understand that transforming raw scores into z scores does not change the shape of the distribution of your scores. If your original data formed a positively skewed distribution, then the distribution of their z scores will remain positively skewed. The same will happen for negatively skewed distributions, or normal distributions. Converting all of the scores to z scores does not affect the shape of the distribution. This is important, for if your data are positively or negatively skewed you should generally not convert the distribution to z scores. However, if the data are normally distributed, then converting the distribution into z scores permits some valuable comparisons and insights.

Let's return to the example of the IQ test. The IQ test is approximately normally distributed with a mean of 100 and a standard deviation of 15. Converting the mean of 100 to a z score is accomplished with the same equation used previously:

$$\text{z score for mean IQ} = \frac{(\mathbf{X} - \mu)}{\sigma}$$

Substituting the known quantities leads to:

$$z = \frac{\mathbf{(100 - 100)}}{\mathbf{15}}$$

This, in turn leads to:

$$\begin{aligned} z &= 0\ /\ 15 \\ &= 0 \end{aligned}$$

Therefore, the mean of the z distribution is 0. This is the case regardless of the variable being considered.

Previously in this chapter it was pointed out that the standard deviation of the common IQ test is 15. And, if we calculated the z score equivalent of an IQ of 115 we would obtain an answer of +1 (you are encouraged to do this calculation). In other words, an IQ of 115 is 1 SD above the mean. In fact, when an entire set of raw scores is converted into z scores the standard deviation will always be converted to 1. Further, as we found in the previous paragraph, the mean of a distribution of z scores will always be 0, regardless of the variable being measured. In other words, *regardless of whether we are dealing with the IQ distribution, or the distribution of points on an* *exam, or any other distribution, when it is converted to z scores, the mean of the new distribution will be 0 and the standard deviation will be 1*. This is very important to remember!

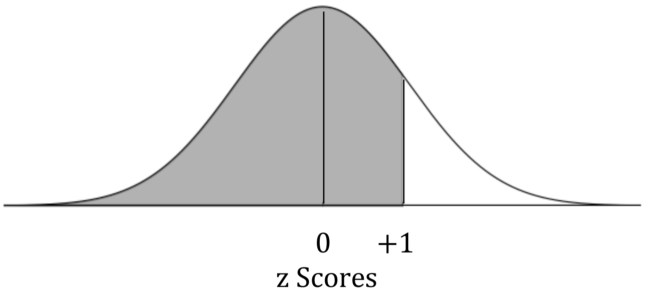
The z Table – Who Would Have Thought That A Few Numbers Could Be So Useful?

We just discussed that with a normal distribution there is a relationship between how many standard deviations a score is from its mean and its precise location on the normal curve. For instance, the inflection points of the normal curve occur at precisely 1 SD from the mean. Further, it was noted that the proportion of the normal curve between the mean and a standard deviation of 1 is always approximately 0.34. We will now see that once *any score's* location on the normal curve is determined, it is possible to specify a series of proportions or probabilities.

We previously noted that the z score associated with an IQ of 70 is –2.0 because an IQ of 70 is equivalent to scoring 2 SD *below* the mean. By referring to the z table (Appendix K, Table 1a for negative values of z) you will see that the entry associated with a z of –2.00 is .02. This is the proportion of the curve below the z score of –2.00 (you are encouraged to draw a figure representing this proportion of the curve). In other words, only 2% of individuals, or 2 people out of 100, would be expected to have an IQ below 70. Of course, then 0.98 is the proportion of the curve that is above our z score of –2.00. Thus, we would expect 98% or 98 people out of 100 to have IQ scores greater than 70. This last proportion is found by subtracting .02, the proportion expected to score below an IQ of 70, from 1.00, the proportion equivalent to the entire distribution.

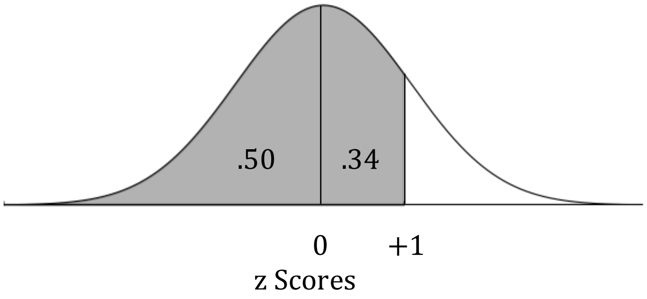
In our IQ example, it was also noted that a test score of 115 is equivalent to a z score of +1.00. By referring to the z table (Appendix K, Table 1b for positive values of z), you will see that the entry associated with a z score of +1.00 is .84, which is the area or proportion of the curve below a z score of +1.00 (Figure 4.7). This indicates that 84% or 84 people out of 100 will score below the z score of +1.00, which corresponds to having an IQ below 115. As you learned in Chapter 2, this percentage, the percentage of scores at or below a particular value, is called the percentile rank.

Figure 4.7 Region Below a z Score of +1



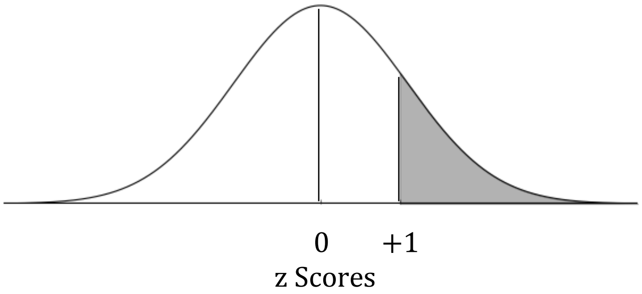
You might wonder why the area of the curve below a z of +1.00 is equal to .84. We know that the normal curve is symmetrical and that the z score of the mean will equal 0. Thus, 50% or .50 of the curve will fall below the mean of 0. And, as we discussed previously, .34 of the distribution will fall between the mean and 1 SD above the mean (a z score of +1.00) (Figure 4.2). Thus, the proportion of the curve below a z score of +1.00 is equal to .50 + .34 which is .84 (Figure 4.8).

Figure 4.8 Determination of the Proportion of the Normal Curve Below a z Score of +1.00



It should also be evident that 1.00 – .84, which is .16, is the proportion of the curve that is above our z score of +1.00 (Figure 4.9). In other words, we would expect 16% or 16 people out of 100 to have IQ scores greater than 115. Clearly, if you are dealing with a normal distribution, once the z score is known, a great deal of additional information is easily obtained.

Figure 4.9 Region Above a z Score of +1



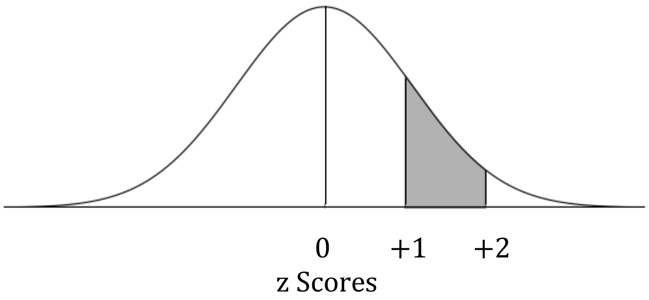
We have limited our discussion to z scores of + or −1, or + or −2. However, it should be evident that by using z tables (Appendix K, Tables 1a and 1b) you can convert any z score into a proportion. For instance, the proportion of the curve falling below a z of −.67 is .25. And if .25 of the distribution falls below a z of −.67, then it follows that .75 of the distribution is above this value of z. Thus you can easily find two proportions associated with any z score so long as you are dealing with a normal distribution. These are the proportion of the curve below the z score, and the proportion of the curve above the z score. Once a proportion is known the percentage is also known, and the number of individuals out of some total can also be determined. In other words, when you are given the raw score, you use the z equation to convert it into a z score and then use the z table to find the corresponding proportion:

raw score > z equation > z score > z table > proportion

And once the proportion is known it can be converted into a percentage, etc.

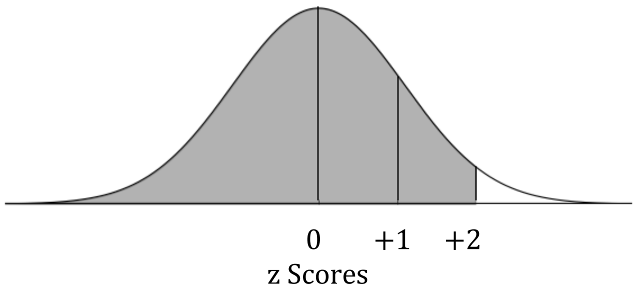
It is also possible to find the proportion or percentage of the distribution that is between two z scores. For instance, what is the proportion of the distribution that is between a z score of +1 and a z score of +2? In order to solve this problem it is best to draw what you are looking for. This is shown in Figure 4.10.

Figure 4.10 Region Between z Scores of +1 and +2



The easiest way to solve this problem is to find the area to the left of a z score of +2 (Figure 4.11), which is .98, and subtract from it the area to the left of a z score of +1 (Figure 4.7), which is .84. The result is .14, the area we are looking for (Figure 4. 10).

Figure 4.11 Region Below a z Score of +2



In the examples we just completed you were given the z scores. Now let's do a complete example, beginning with the raw scores. The scores on the SAT exam are approximately normally distributed and have a mean of 500 and a standard deviation of 100. How many people, out of 1000, would be expected to score between 350 and 575?

The first step is to convert the two raw scores into z scores. This is shown below:

$$z = \frac{(\mathbf{X} - \mu)}{\sigma}$$

$$\text{z for an SAT score of } 350 = \frac{\mathbf{(350 - 500)}}{\mathbf{100}}$$

$$= \frac{\mathbf{-150}}{\mathbf{100}}$$

$$= -1.50$$

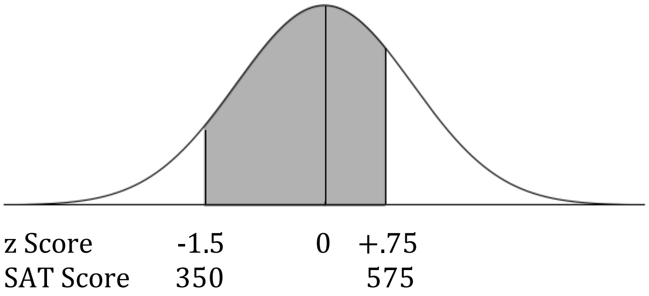
$$\text{z for an SAT score of } 575 = \frac{\mathbf{(575 - 500)}}{\mathbf{100}}$$

$$= \frac{\mathbf{75}}{\mathbf{100}}$$

$$= +.75$$

We now can draw the region of the normal curve, using SAT scores as well as z scores, that is of interest to us. This is shown in Figure 4.12.

Figure 4.12 Region Between SAT Scores of 350 and 575



Once the z scores that are equivalent to SAT scores of 350 and 575 have been calculated, the easiest way to find the desired region is to find the proportion of the curve that is below a z score of +.75 and subtract from this the proportion of the curve that lies below a z score of −1.50.

By referring to the z table (Appendix K, Table 1b), you will see that the entry associated with a z score of +.75 is .77 which, you recall, is the proportion of the curve below this z score. From this we subtract the proportion of the curve which is below a z score of −1.50, which is .07 (Appendix K, Table 1a). The result, .70, is the proportion of the curve that falls between z scores of −1.50 and +.75, the z scores equivalent to SAT scores of 350 and 575.

While essential to finding the answer, this is not what the question asked us to find. The problem was to find *how many people, out of 1000*, would be expected to have SAT scores between 350 and 575. To find this we must multiply our proportion of .70 by 1000, the total number of people. This results in 700 people out of 1000 being expected to have SAT scores between 350 and 575.

Progress Check

1. The magnitude of a z score indicates the number of _____ that a score falls from the mean, and the _____ indicates whether the score is larger or smaller than the mean.

Use the following information for the next two problems: SAT exam scores are normally distributed. The mean is 500 and the standard deviation is 100.

2. What is the z score that is equivalent to an SAT score of 700?

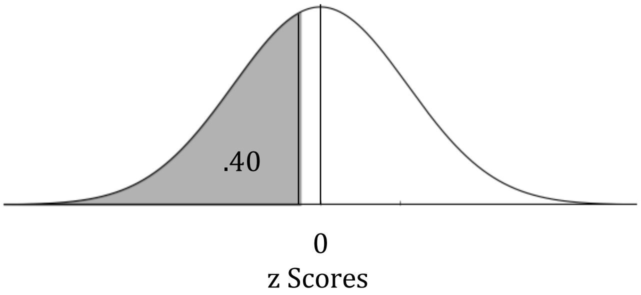
3. What proportion of SAT scores would fall below an SAT score of 350?

Answers: 1. standard deviations; sign 2. +2.00 3. .07

We will conclude this chapter by discussing two additional types of problems that utilize z scores. The first deals with finding the raw score that is associated with a particular proportion of the curve. We know that a z score of 0 is at the mean of a distribution and, since the normal distribution is symmetrical, half of the area is below the mean and half the area is above the mean. But what about other proportions, such as .40? What z score has .40 of the curve below it? In other words, what z score has a percentile rank of 40%?

The solution is found by first noting that if 40% of the distribution falls below our z score, then we are to the left of the mean since 50% of the distribution falls below the mean of a symmetrical distribution. The region of the distribution that we are interested in is illustrated in Figure 4.13. And we know that our z score will be negative. We now refer to the body of the z table (Appendix K, Table 1a) and look for the proportion .40. The proportion .40 occurs twice in the z table, and is associated with a z score of –.25 or –.26. (Either value is accurate enough for our purposes. We will choose –.25 for our calculations.) In other words, approximately .40, or 40%, of the distribution occurs below a z score of –.25. Alternatively, we could say that the percentile rank of a z score of –.25 is 40%.

Figure 4.13 Region with a Percentile Rank of .40 (40%)



However, what if the problem did not ask for the z score, but instead it asked for the equivalent raw score? This requires some additional calculation, but is not difficult. For instance, using the example of SAT scores, it is easy to convert the z score with a percentile rank of 40% (which we just determined was –.25) into an SAT score. To do this we use either the definitional equation for the z score or the rearranged equation that was illustrated earlier. The solution, using both forms of the equation, is presented below:

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

$$-.25 = \frac{(\mathbf{X - 500})}{\mathbf{100}}$$

$$-25 = X - 500$$

$$475 = X$$

or

$$X = z\sigma + \mu$$

$$= [(-.25)(100)] + 500$$

$$= -25 + 500$$

$$= 475$$

Thus, we have found that a percentile rank of 40% is equivalent to a z score of **–**.25, which in turn is equivalent to an SAT score of 475. In other words, when you are given the percentile rank you convert it into a z score using the z table and then use the z equation to find the corresponding raw score:

percentile rank > z table > z score > z equation > raw score

For our second example, what IQ score would have 80% of the population above it? Since our z table only gives the proportions below a z score you need to recognize that this is equivalent to asking what score would have 20% or .20 of the distribution below it. (I suggest you drawn this.) We now need to convert the proportion .20 into a z score and, since this z score will fall to the left of the mean, it will be found to be negative. Next we refer to the body of the z table (Appendix K, Table 1a) and look for the proportion closest to .20. The value of .20 occurs four times in the table. You could turn to a table with more precise values of z or, alternatively, pick the middle value of z from our table. This would be **–**.84 or **–**.85. Either value is sufficiently accurate for us. I have chosen **–**.84 for the calculations. To find the IQ score that is equivalent to a z score of **–**.84 recall that the mean of an IQ test is 100 and the standard deviation is 15. We then substitute into the definitional equation (or the rearranged equation given previously):

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

$$-.84 = \frac{(\mathbf{X - 100})}{\mathbf{15}}$$

$$-12.6 = X - 100$$

$$87.4 = X$$

or

$X = z\sigma + \mu$

$= [(-.84)(15)] + 100$

$= -12.6 + 100$

$= 87.4$

To summarize our steps, we first recognized that an IQ score with 80% of the population above it must have 20% of the population below it. And 20% is equivalent to .20, which corresponds to a z score of –.84. Finally, we found a z score of –.84 is equivalent to an IQ score of 87.4.

Conclusion

This chapter focused on the z score, which is used with interval or ratio data. The z score is known as a standard score, and is defined as the number of standard deviations that a score differs from its mean. The equation for the z score is, therefore, $z = (X - \mu) / \sigma$.

We found that converting a distribution into z scores results in a distribution with a mean of 0 and a standard deviation of 1. However, this conversion does not change the distribution's shape. If the distribution was skewed originally, it remains skewed. If the original distribution was normal, then it remains normal. Once normally distributed scores are converted to z scores it is possible, using the z table (Appendix K, Tables 1a and 1b), to ascertain the proportions of the distribution that are associated with any particular z score. This is valuable in answering a variety of questions about IQ, SAT, or other normally distributed sets of scores. For instance, we learned that we could easily find the proportion of the curve located between two scores, as well as the percentile rank of a score. In addition, by converting raw scores into z scores it is possible to compare outcomes measured on different scales. For instance, it is possible with z scores to compare outcomes on a 10-point quiz and a 100-point exam, even though the measurement scales differ dramatically. It is also possible to convert a z score back to a raw score.

It should come as no surprise that the z score is a commonly used descriptive measure of variability.

Glossary Of Terms

Inflection point – A point on a graph where the curvature changes from concave to convex or from convex to concave.

Raw score – Your data as they are originally measured, before any transformation.

<u>Standard score</u> – A measure indicating whether a score is above or below the mean as well as how many standard deviations it is from the mean. Also called a z score.

<u>z score</u> – Conversion of raw data so that the deviation is measured in standard deviation units and the sign, positive or negative, indicates the direction of the deviation.

Questions – Chapter 4

(Answers are provided in Appendix J.)

1. A z score of –2 indicates that the point is _____.
   a. 2 standard deviations below the mean
   b. 2 standard deviations above the mean
   c. twice the mean
   d. half the mean

2. A 'standard' normal curve has a mean of _____ and a standard deviation of _____.
   a. 1; 0
   b. 100; 10
   c. 0; 1
   d. 10; 100

3. In a normal curve, the inflection points occur at _____ standard deviation(s) from the mean.
   a. +/ –10
   b. +/ –1
   c. 0
   d. depends upon the specific curve

4. Another name for the z score is the _____ score.
   a. normal
   b. special
   c. independent
   d. standard

5. If a distribution of scores is positively skewed, converting each score into a z score will result in a distribution which is _____.
   a. positively skewed
   b. negatively skewed
   c. normal
   d. cannot be answered without additional information

6. What percent of scores fall below an IQ of 85?
   a. 8
   b. 10
   c. 16
   d. 25

7. On the IQ test, what percent of people score below 90?
   a. 14
   b. 16

c. 21
d. 25

8. On the IQ test, what percent of people score between 90 and 130?
a. 82
b. 73
c. 61
d. 50

9. On the IQ test, what percent of people score above 110?
a. 25
b. 27
c. 29
d. 31

10. Out of a population of 1000 individuals, how many would you expect to have an IQ greater than 85?
a. 670
b. 734
c. 803
d. 840

11. Assuming a normally distributed population with a mean of 50 and a standard deviation of 5, how many people, out of 100, would you expect to score higher than 58 or lower than 48?
a. 30
b. 39
c. 52
d. 66

12. What IQ score results in 20% of the population scoring above it?
a. 100
b. 130
c. 112.6
d. 119.2

13. What score results in 40% of the population scoring below it, assuming a mean of 25 and a standard deviation of 4?
a. 21
b. 22
c. 23
d. 24

14. What score results in 65% of the population scoring below it, assuming a mean of 10 and a standard deviation of 5?
a. 11.9
b. 10.5
c. 11.2
d. 12.7

15. On any normal distribution, the 50th percentile corresponds to a z score of _____.
a. 0
b. +2
c. +1
d. –1

16. A z score of +2 indicates that the raw score is _____.
    a. 2 standard deviations below the mean
    b. 2 percentage points below the mean
    c. 2 standard deviations above the mean
    d. 2 percentage points above the mean

17. With a normal curve, the probability of a score occurring <u>above</u> the mean is _____.
    a. 0
    b. 0.5
    c.. 75
    d. Cannot be determined

18. On an exam, a student would prefer their outcome to be equivalent to a z score of _____.
    a. –1
    b. +1
    c. +0.25
    d. 0

19. With a normal distribution, how many people, out of 100, would you expect to score *between* –1 and +2 standard deviations from the mean?
    a. 82
    b. 16
    c. 66
    d. 84

20. The SAT exam has a mean of 500 and a standard deviation of 100. What is the z score for an SAT exam score of 415?
    a. +.15
    b. –.15
    c. +.85
    d. –.85

Chapter 5
Using IBM SPSS Statistics 26

"The saddest aspect of life right now is that science gathers knowledge faster than society gathers wisdom."

Isaac Asimov

Introduction

The calculations involved in solving statistical problems can become tedious, particularly if there is a large set of data. The initial response in the field of statistics was to derive a series of what are called **Computational Equations**. These equations made analyzing large data sets somewhat easier, however this text relegates them to the appendixes. This is because during the past few decades the availability of computers and powerful statistical software has become commonplace. The result has been a revolution in how statistical problems are actually solved. Though it is still important for statistics students to learn how to calculate answers, preferably using definitional equations so you understand what you are calculating, it is also important for students to learn to use a statistical software package. One of the most widely used of these software packages is called **SPSS**.

Computational equations – Equations developed to aid in statistical calculations. They were useful with large data sets, but now researchers would employ computer software packages instead.

SPSS – A powerful, commonly-used statistical computer package. The letters 'SPSS' originally were an abbreviation for 'statistical package for the social sciences'.

SPSS has undergone numerous revisions. The result is a flexible, user-friendly program. In this chapter you will be introduced to this very important tool. More specifically, we will be utilizing examples from previous chapters so you can become acquainted with labeling variables, entering data and creating bar graphs, pie charts and histograms. In subsequent chapters you will learn additional features of SPSS. Do not be concerned if you are not a computer expert, basic SPSS is very easy to master.

SPSS is organized so that each column is a variable and each row consists of a subject's data (Figure 5.1). There can be as many variables (columns) or subjects (rows) as desired. You do not need to worry about having too many of either.

Figure 5.1 Organization of Data in SPSS

Columns

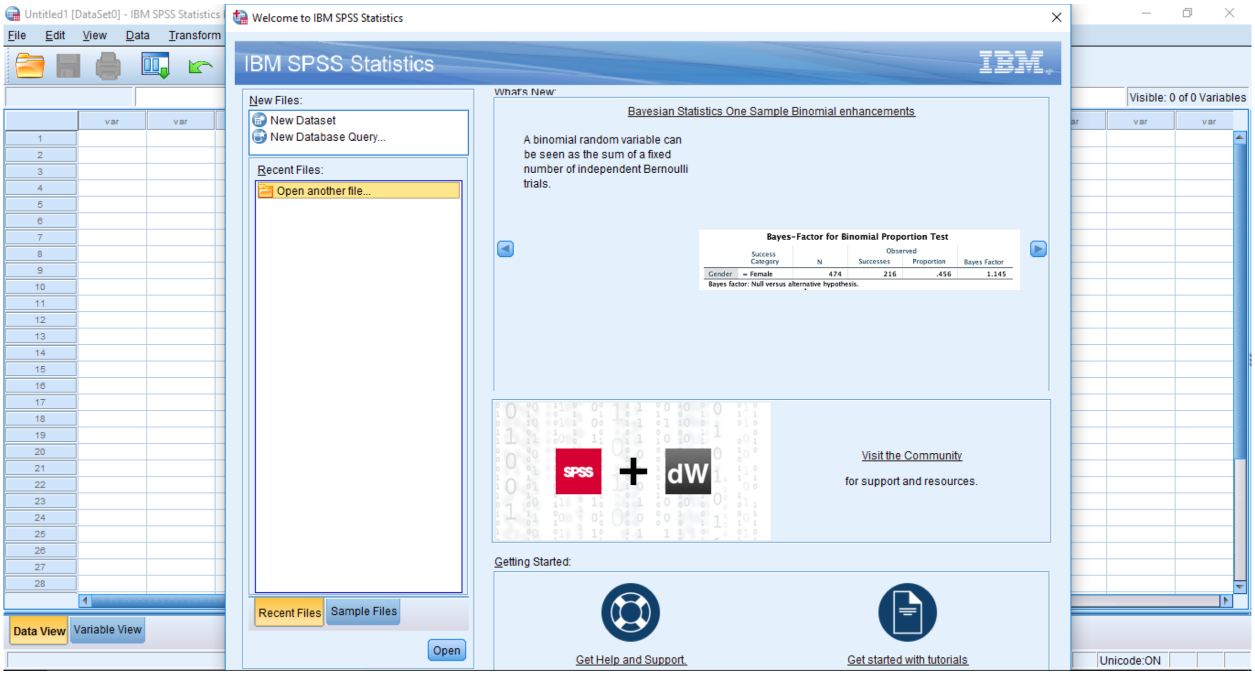
		Variable 1	Variable 2	Variable 3	Variable 4	Etc.
	Subject 1					
Rows	Subject 2					
	Subject 3					
	Etc.					

Our First Example (A) Using SPSS

To Begin SPSS

Step A.1 The first step is to activate the program. You can accomplish this by double clicking on the SPSS icon on the computer's desktop or, if this icon is not evident, by clicking on 'Program' and then on SPSS. You will see the window displayed in Figure 5.2. (This text uses SPSS version 26. Other versions of SPSS will have a very similar window.) At this point you have a number of options. If you click on the 'Get started with tutorials' you will be guided through a very informative, general introduction to using SPSS. The current chapter's goal is more limited – you will learn how to define variables, enter data by hand and conduct the descriptive statistics that you have learned in Chapters 1 – 4. Therefore, click on the 'X' at the top right of the central window, or the 'Close' button at the bottom right of the central window. (Unfortunately, the 'Close' button is cut off in Figure 5.2.)

Figure 5.2 The Initial SPSS Window



Bayesian Statistics One Sample Binomial enhancements

A binomial random variable can be seen as the sum of a fixed number of independent Bernoulli trials.

Bayes-Factor for Binomial Proportion Test

	Success Category	N	Observed		Bayes Factor
			Successes	Proportion	
Gender	= Female	474	216	.456	1.145

Bayes factor: Null versus alternative hypothesis.

Visit the Community

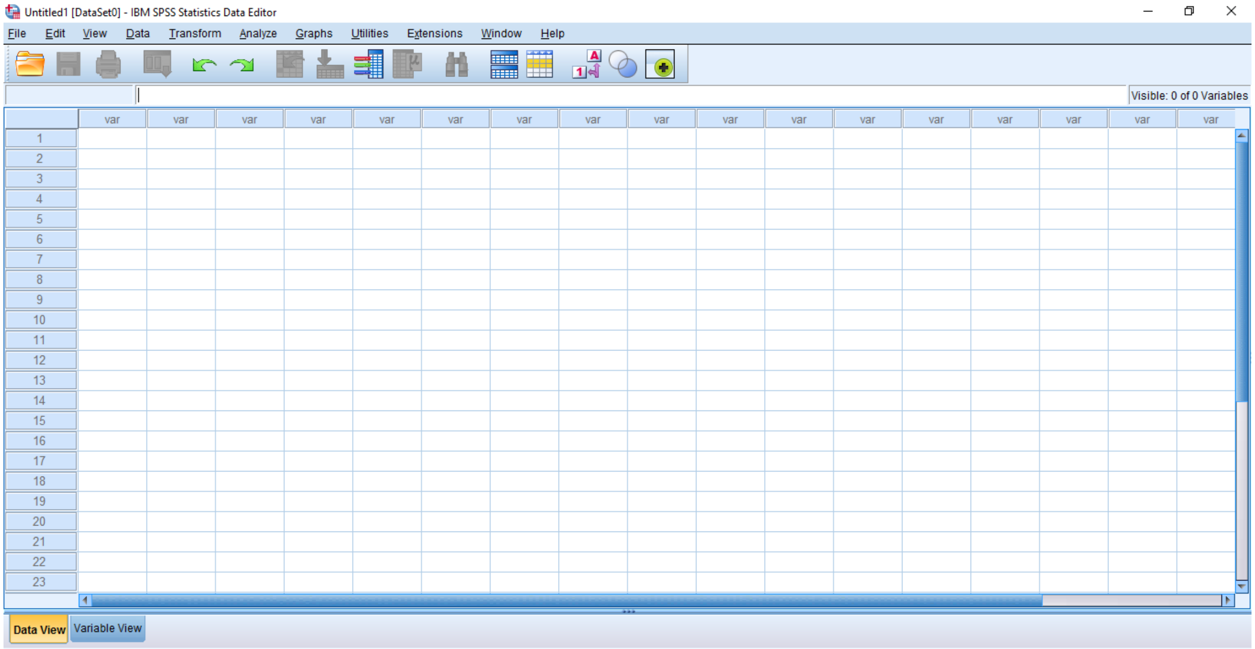
for support and resources.

Step A.2 The blank screen shown in Figure 5.3 appears. SPSS utilizes two windows. At the bottom left of the current window are two 'switches', one labeled **Data View**, the other **Variable View**. As the name suggests, the Data View window shows the data that the SPSS program is currently using. The Variable View window provides information concerning the variables that are listed in the Data View window.

Data view – SPSS window in which the data are displayed.

Variable view – SPSS window in which variables are defined.

Figure 5.3 The Data View Window



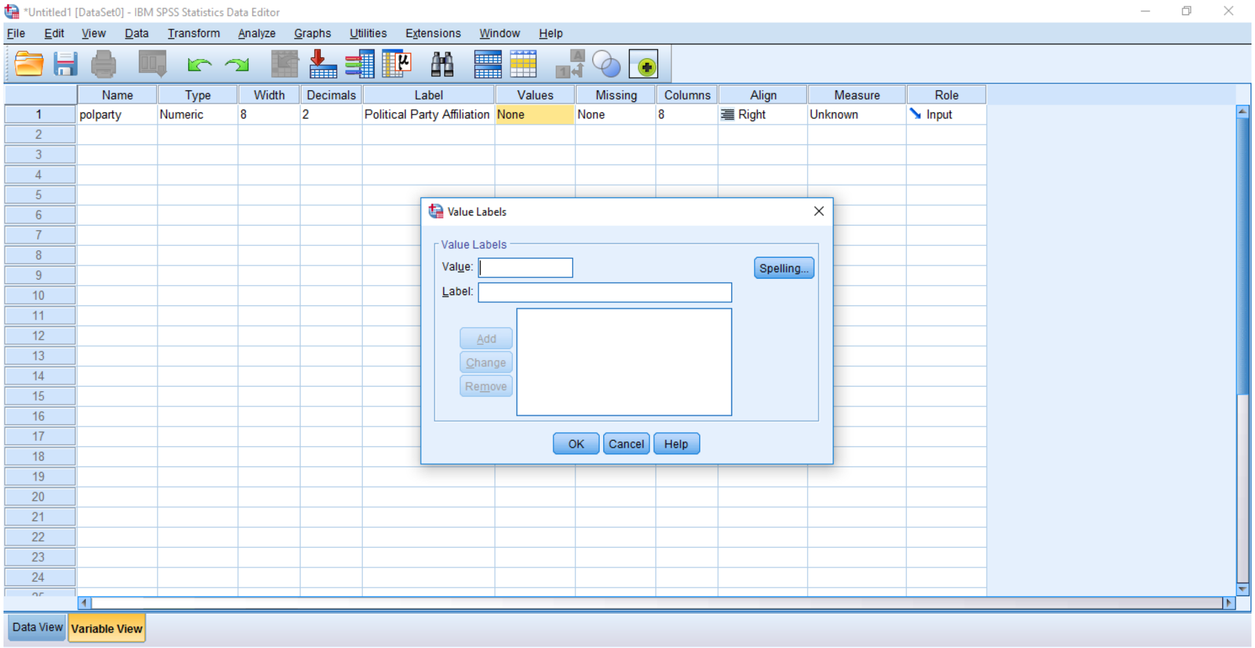
Step A.3 Click on 'Variable View'. This brings up a window that superficially looks like the Data View window (you can switch back and forth between them). Near the top of this page is a row of column headings, beginning with 'Name', then 'Type', and proceeding to 'Role'. For the present we will only be dealing with the columns headed by 'Name', 'Label', 'Values' and 'Measure'.

Step A.4 Click on the first empty rectangle (called a 'cell') under the column heading 'Name'. The upper left 'cell' will turn yellow. You now type the name of the first variable for which you have data. We are going to utilize the same data and labels as were previously employed in Table 2.3. As these data dealt with the political preferences of a group of hypothetical subjects we have only one variable which I call 'polparty'.

Step A.5 Click on the first empty 'cell' under the column heading 'Label'. In this cell you can type a more extensive description of your variable. In our case, type 'Political Party Affiliation'. Note that in order to see the entire label you may need to expand the size of this cell by placing your cursor on the right border of the Label heading and moving to the right.

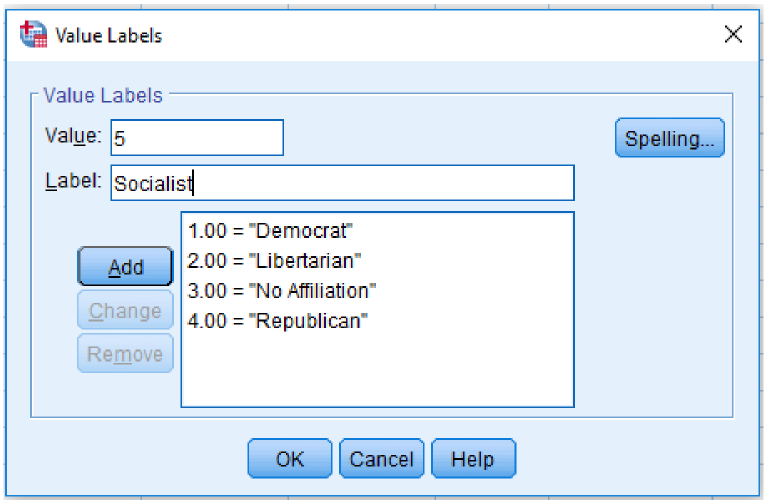
Step A.6 Click on the first empty 'cell' under the column heading 'Values'. A box will appear as in Figure 5.4. For most analyses SPSS utilizes numbers. Thus, we will need to assign a number for each political party affiliation. In the blank space to the right of 'Value', type the number '1'. Then type a brief description of this value of the variable in the blank space to the right of 'Label'. In our case, type 'Democrat'. Finally, click on 'Add'. Your label for a value of 1 will appear in the large white region in the center of the window. Now repeat the above steps in this section for each of the values in the data set as listed in Table 2.4. Figure 5.5 illustrates what you will see immediately before clicking on 'Add' after defining all of the values of the data set. After clicking on 'Add', then click on 'OK'.

Figure 5.4 The Value Labels Box of the Variable View Window



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	polparty	Numeric	8	2	Political Party Affiliation	None	None	8	Right	Unknown	Input

Figure 5.5 The Assignment of Value Labels



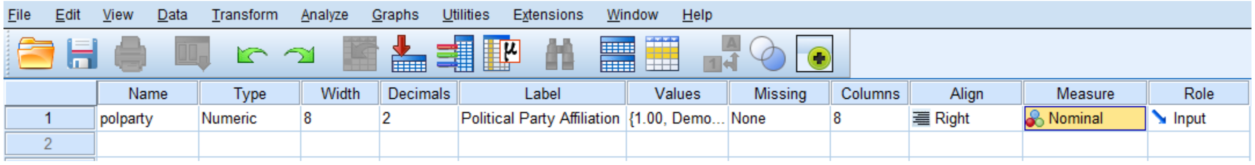
Value: 5

Label: Socialist

1.00 = "Democrat"
2.00 = "Libertarian"
3.00 = "No Affiliation"
4.00 = "Republican"

Step A.7 Click on the first empty 'cell' under the column heading 'Measure'. As we are dealing with nominal data, select 'Nominal' as is shown in Figure 5.6. You have now completed the 'Variable View' window for the data we are interested in.

Figure 5.6 The Completed Variable View Window

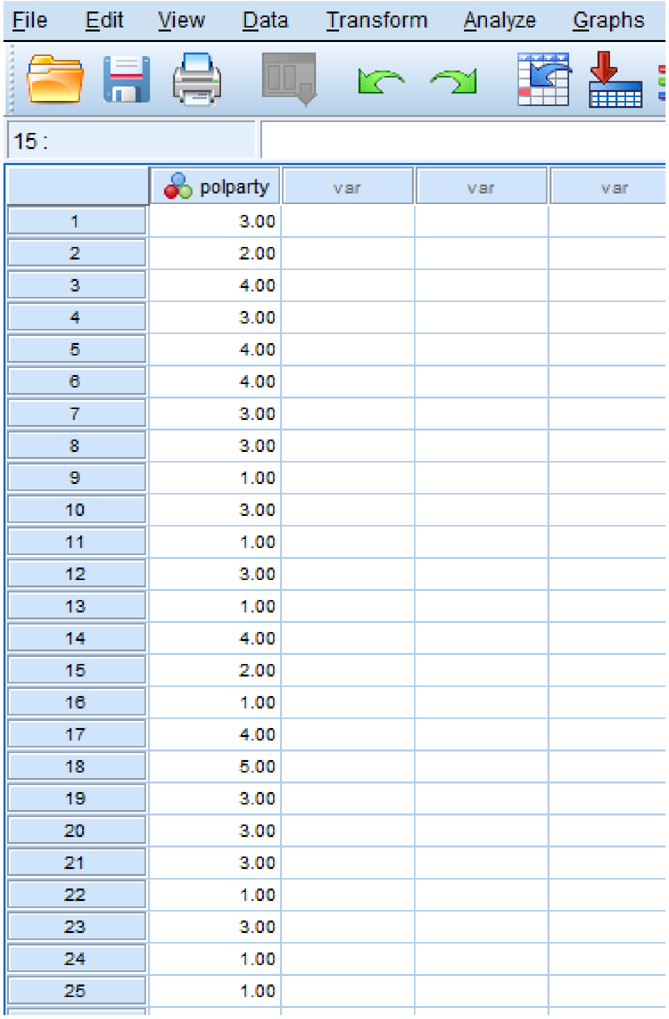


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	polparty	Numeric	8	2	Political Party Affiliation	{1.00, Demo...	None	8	Right	Nominal	Input
2											

Step A.8 Click on the 'Data View' option at the lower left corner of the window. The label 'polparty' will be present in the first variable column.

Step A.9 Click on the first empty 'cell' under 'polparty' and, while referring to Table 2.3, type in the number corresponding to the political affiliation of the first subject, in this case '3' as they did not indicate a party affiliation. Continue adding data by clicking on the next empty 'cell' in the column under 'polparty' until all 25 values have been entered (the first 23 values are shown in Figure 5.7). Note that I entered the data by going down the columns in Table 2.3, but entering the data by going across the rows would have worked just as well. You are now ready to use SPSS to describe the data you have entered.

Figure 5.7 Data View Window

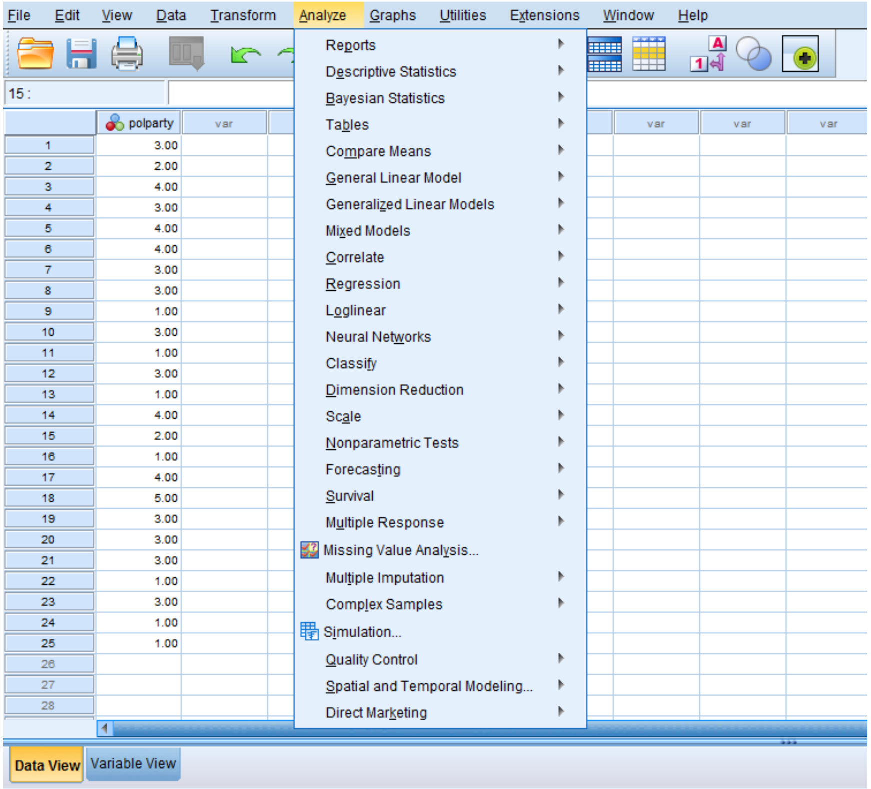


	polparty	var	var	var
1	3.00			
2	2.00			
3	4.00			
4	3.00			
5	4.00			
6	4.00			
7	3.00			
8	3.00			
9	1.00			
10	3.00			
11	1.00			
12	3.00			
13	1.00			
14	4.00			
15	2.00			
16	1.00			
17	4.00			
18	5.00			
19	3.00			
20	3.00			
21	3.00			
22	1.00			
23	3.00			
24	1.00			
25	1.00			

To Find Frequencies

Step A.10 Click on '**Analyze**' at the top of the window as is shown in Figure 5.8.

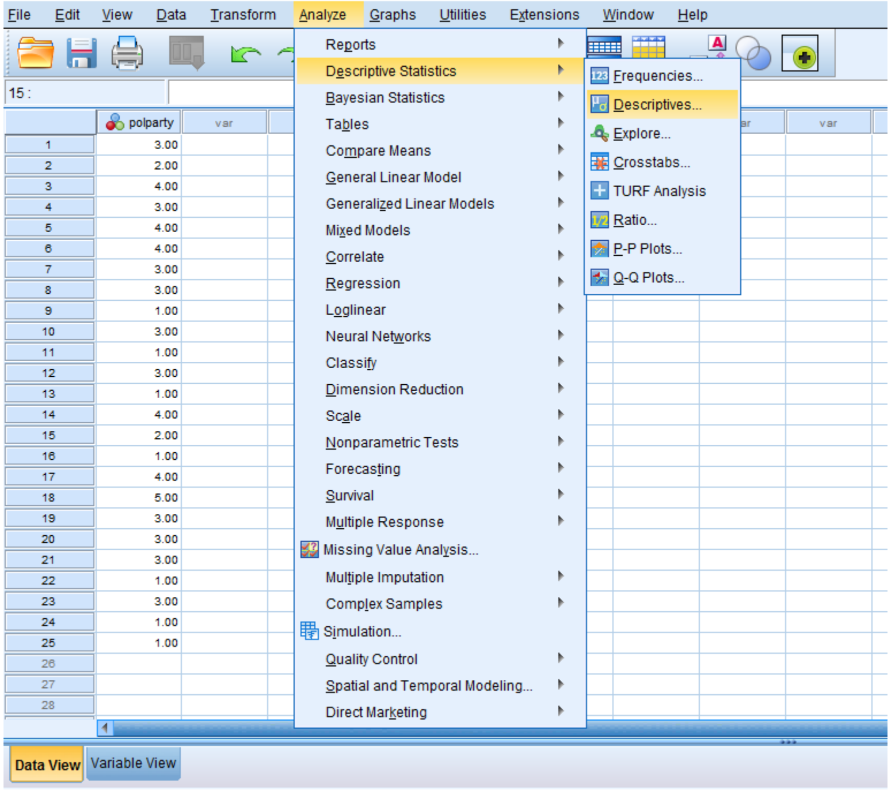
Figure 5.8 The Analyze Function



	polparty
1	3.00
2	2.00
3	4.00
4	3.00
5	4.00
6	4.00
7	3.00
8	3.00
9	1.00
10	3.00
11	1.00
12	3.00
13	1.00
14	4.00
15	2.00
16	1.00
17	4.00
18	5.00
19	3.00
20	3.00
21	3.00
22	1.00
23	3.00
24	1.00
25	1.00
26	
27	
28	

Step A.11 Move your cursor down to '**Descriptive Statistics**'. When you do so, an additional window will appear (Figure 5.9).

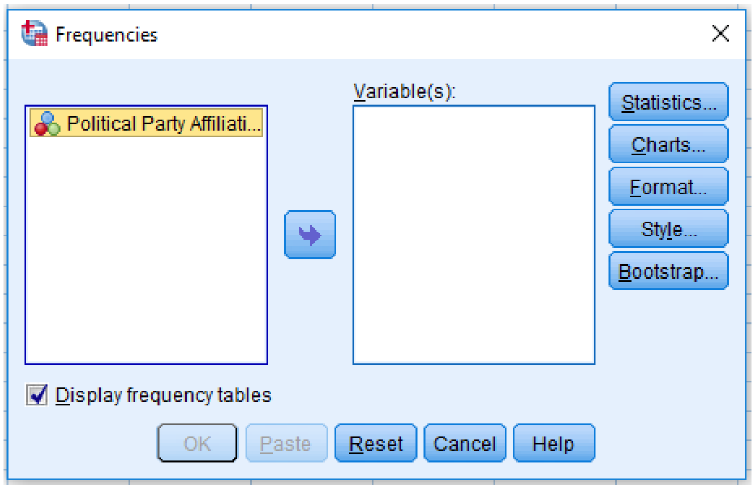
Figure 5.9 The Descriptive Statistics Function



	polparty	var
1	3.00	
2	2.00	
3	4.00	
4	3.00	
5	4.00	
6	4.00	
7	3.00	
8	3.00	
9	1.00	
10	3.00	
11	1.00	
12	3.00	
13	1.00	
14	4.00	
15	2.00	
16	1.00	
17	4.00	
18	5.00	
19	3.00	
20	3.00	
21	3.00	
22	1.00	
23	3.00	
24	1.00	
25	1.00	
26		
27		
28		

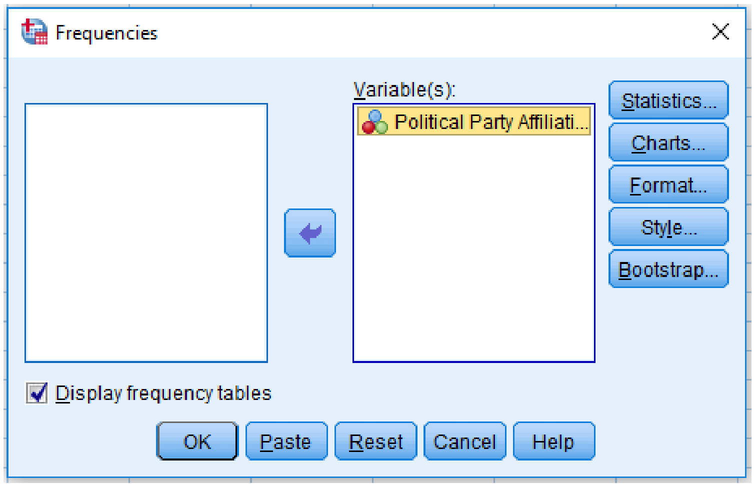
Step A.12 Click on ‘**Frequencies**’ and the window shown in Figure 5.10 will appear.

Figure 5.10 The Frequencies Window



Step A.13 Click on the arrow symbol pointing to the right to move the variable label, ‘Political Party Affiliation’ to the rectangle with the heading ‘Variable(s)’ (Figure 5.11). (As we only have one variable this may seem un-necessary, but when you have numerous variables this is how to indicate to SPSS which variables are currently of interest.)

Figure 5.11 Using the Frequencies Window

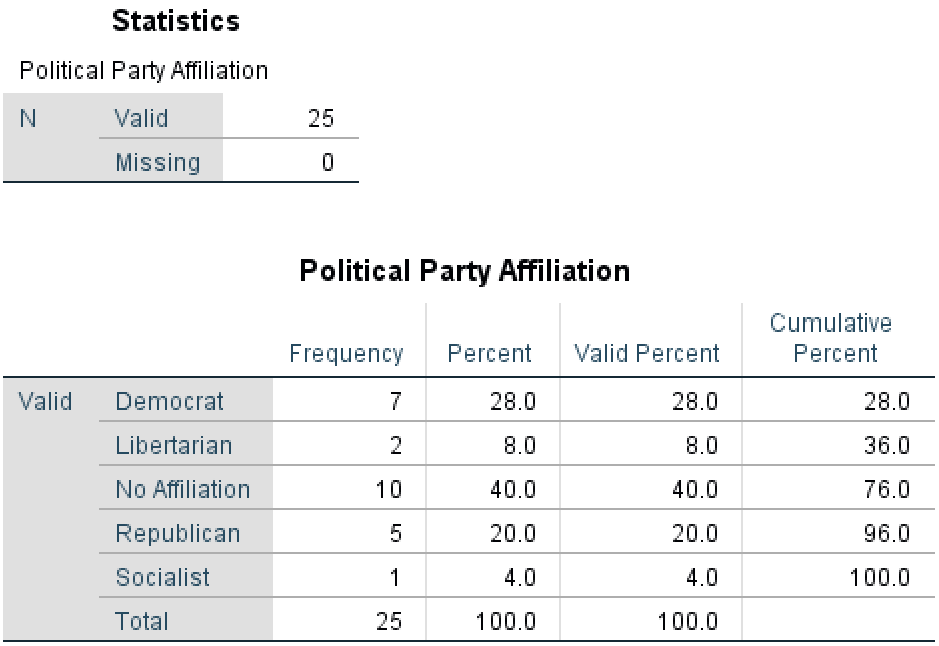


Variable(s): Political Party Affiliati...

☑ Display frequency tables

Step A.14 Click on '**OK**'. The result is the output shown in Figure 5.12, which includes the frequencies previously shown in Table 2.4 as well as additional information that may be of interest. This output can be printed and/or saved, if desired.

Figure 5.12 An Example of SPSS Output



Statistics

Political Party Affiliation

N	Valid	25
	Missing	0

Political Party Affiliation

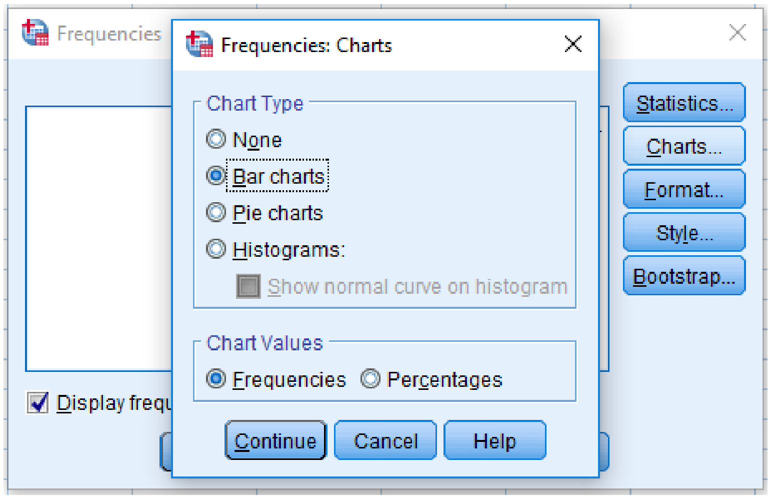
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Democrat	7	28.0	28.0	28.0
	Libertarian	2	8.0	8.0	36.0
	No Affiliation	10	40.0	40.0	76.0
	Republican	5	20.0	20.0	96.0
	Socialist	1	4.0	4.0	100.0
	Total	25	100.0	100.0	

Step A.15 Exit from this output (not from SPSS). You will be prompted whether you want to save the output. As this is just an exercise, you do not have to save it.

To Create A Bar Graph

Step A.16 You will have been returned to the 'Data View' window. Once again, click on '**Analyze**' at the top of the window, move your cursor down to '**Descriptive Statistics**', and then click on '**Frequencies**'. You should return to the window shown in Figure 5.11. Click your cursor on '**Charts**', and indicate which of the charts or graphs is wanted. As we desire a bar graph, '**Bar charts**' has been selected in Figure 5.13.

Figure 5.13 Selecting a Bar Graph



Frequencies: Charts

Chart Type

○ None
● Bar charts
○ Pie charts
○ Histograms:
☐ Show normal curve on histogram

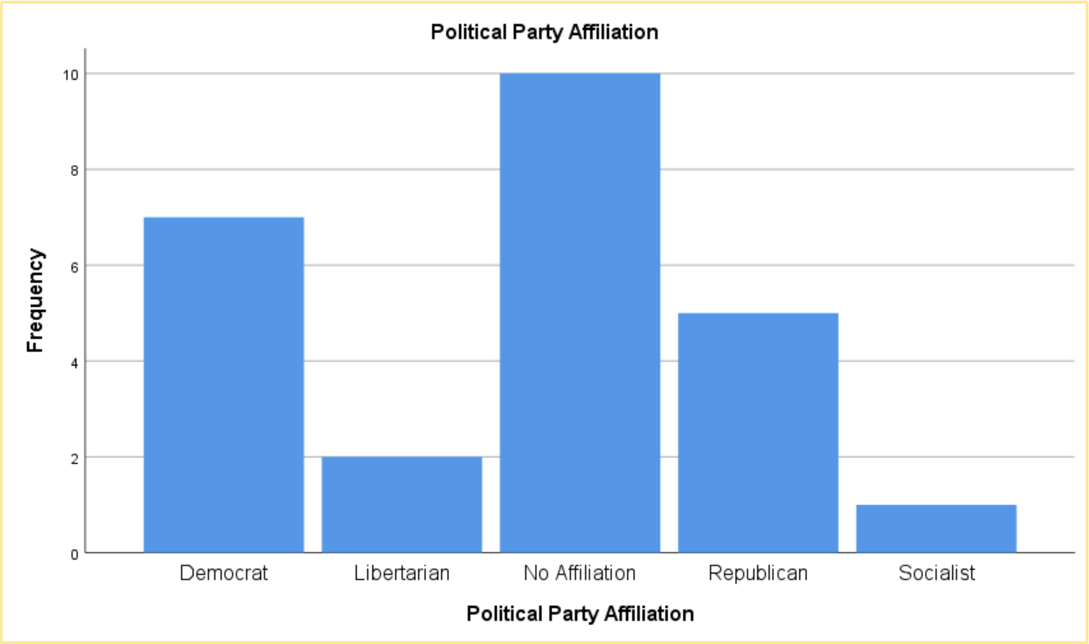
Chart Values

● Frequencies ○ Percentages

☑ Display frequ

Step A.17 Click on '**Continue**'. If you do not want to see the frequency distribution again, click on the check in front of the phrase '**Display frequency tables**', and then '**OK**' and you will see just the bar graph shown in Figure 5.14.

Figure 5.14 The Bar Graph



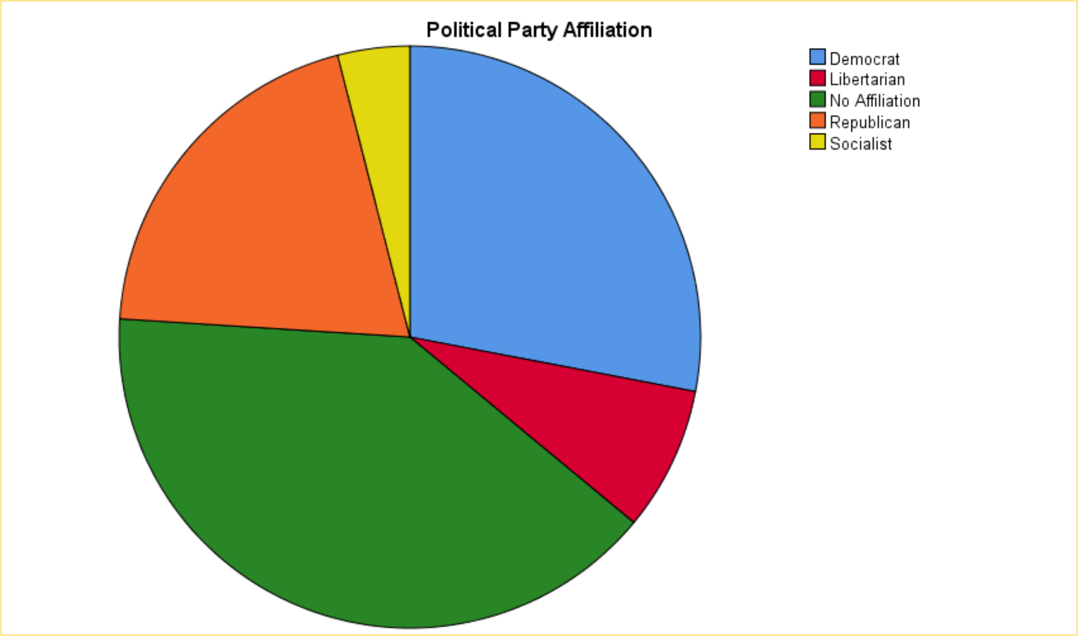
Step A.18 Exit from this output (not from SPSS) by clicking on the X in the upper right corner. You will be prompted whether you want to save the output. As this is just an exercise, you do not have to save this graph.

To Create A Pie Chart

Step A.19 To create a pie chart of your frequencies, we repeat the steps used for creating a bar graph except that when we reach Figure 5.13 we select '**Pie charts**'.

Step A.20 Click on '**Continue**'. If you do not want to see the frequency distribution again, click on the check in front of the phrase '**Display frequency tables**', and then click '**OK**' and you will see just the pie chart shown in Figure 5.15.

Figure 5.15 The Pie Chart



Step A.21 Exit from this output by clicking on the X in the upper right corner. You will be prompted whether you want to save the output. Once again, as this is just an exercise, you do not have to save this chart. If you continue exiting you will return to the Data View window. Exiting from this window will result in another prompt asking if you want to save the data. As this is just an example, there is no need to save the data unless you feel you might want to practice making bar graphs and pie charts using these data in the future.

A Second Example (B) Using SPSS

To Begin SPSS Proceed As In The Previous Example:

Step B.1 This is the same as step A.1.

Step B.2 This is the same as step A.2.

Step B.3 This is the same as step A.3.

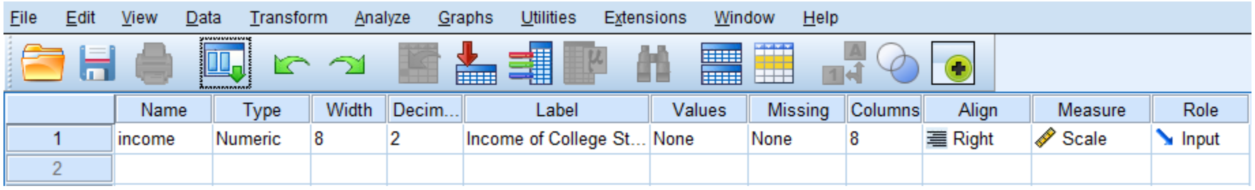
Step B.4 Click on the first empty cell under the column heading ‘Name’. We are going to utilize the same data and labels as were previously employed in Table 3.6. As these data dealt with the incomes of 10 hypothetical students, we have only one variable which I call ‘income’.

Step B.5 Click on the first empty ‘cell’ under the column heading ‘Label’. In this cell I have typed ‘Income of College Students’ as a more extensive description of our variable. Note that in order to see the entire label you will need to expand the size of this cell by placing your cursor on the right border of the Label heading and moving to the right.

Step B.6 In the current example all of the data are expressed in dollars. There is not, therefore, any need to label ‘Values’ of your variable.

Step B.7 Under the column heading ‘Measure’, choose ‘Scale’. In SPSS, ‘Scale’ indicates that the data are at either the interval or ratio level of measurement. As our data are measured in dollars, and this is a ratio level of measurement, ‘Scale’ is the appropriate entry. You have now completed the ‘Variable View’ window for the data we are interested in (Figure 5.16).

Figure 5.16 Defining a Label Within the Variable View Window

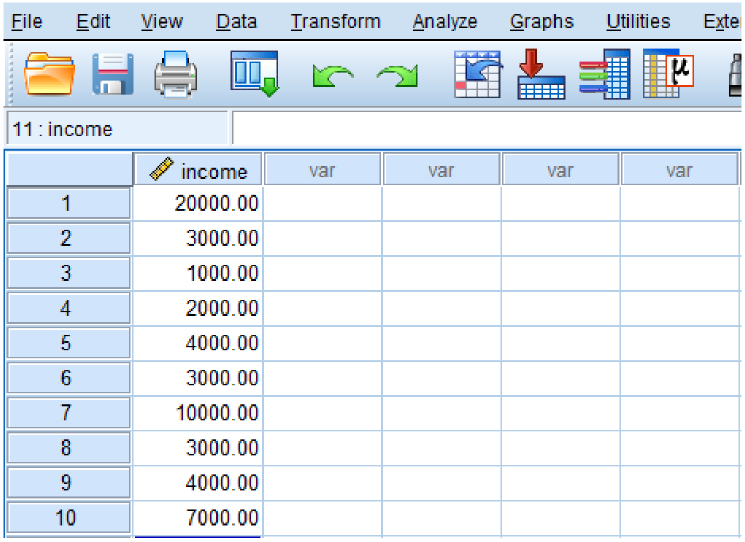


	Name	Type	Width	Decim...	Label	Values	Missing	Columns	Align	Measure	Role
1	income	Numeric	8	2	Income of College St...	None	None	8	Right	Scale	Input
2											

Step B.8 Click on the ‘Data View’ option at the lower left corner of the window. The variable ‘income’ will now be present.

Step B.9 Click on the first empty ‘cell’ under ‘income’ and type in the number corresponding to the income of the first student, in this case 20000. Continue adding data by clicking on the next empty ‘cell’ in the column under ‘income’ until all 10 values have been entered (Figure 5.17). You are now ready to use SPSS to describe the data you have entered.

Figure 5.17 Entering Data in the Data View Window



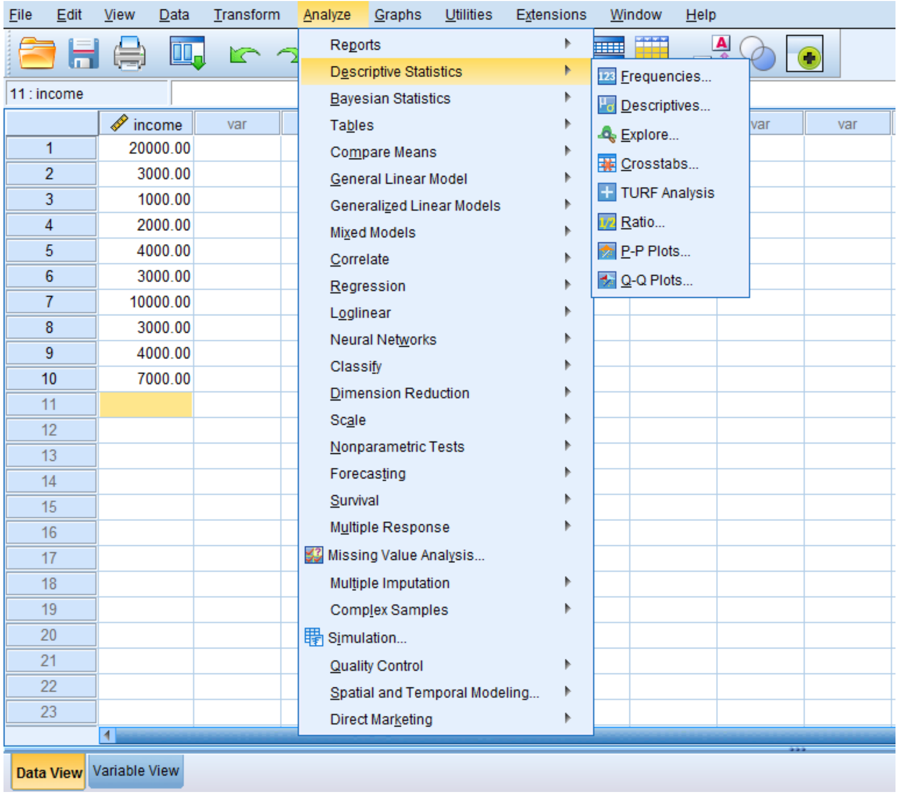
	income	var	var	var	var
1	20000.00				
2	3000.00				
3	1000.00				
4	2000.00				
5	4000.00				
6	3000.00				
7	10000.00				
8	3000.00				
9	4000.00				
10	7000.00				

To Find Frequencies

Step B.10 Click on ‘**Analyze**’ at the top of the Data View window.

Step B.11 Move your cursor down to ‘**Descriptive Statistics**’. When you do so, an additional window will appear (Figure 5.18).

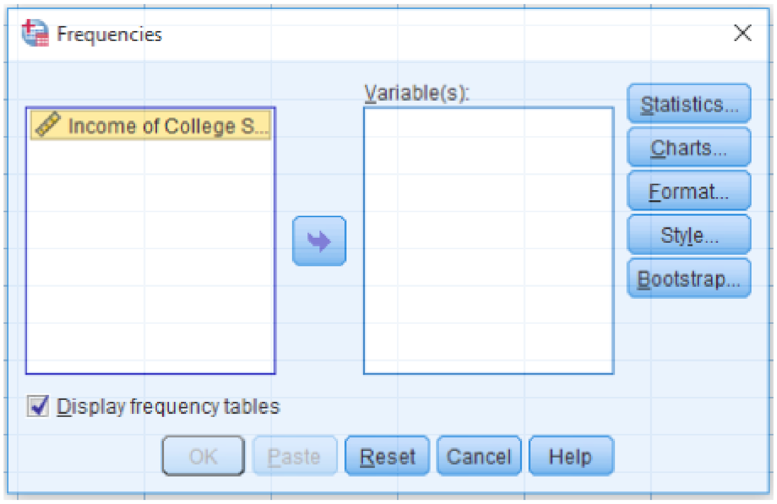
Figure 5.18 The Descriptive Statistics Function



	income	var
1	20000.00	
2	3000.00	
3	1000.00	
4	2000.00	
5	4000.00	
6	3000.00	
7	10000.00	
8	3000.00	
9	4000.00	
10	7000.00	
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		
21		
22		
23		

Step B.12 Click on '**Frequencies**' and the window shown in Figure 5.19 will appear.

Figure 5.19 The Frequencies Window

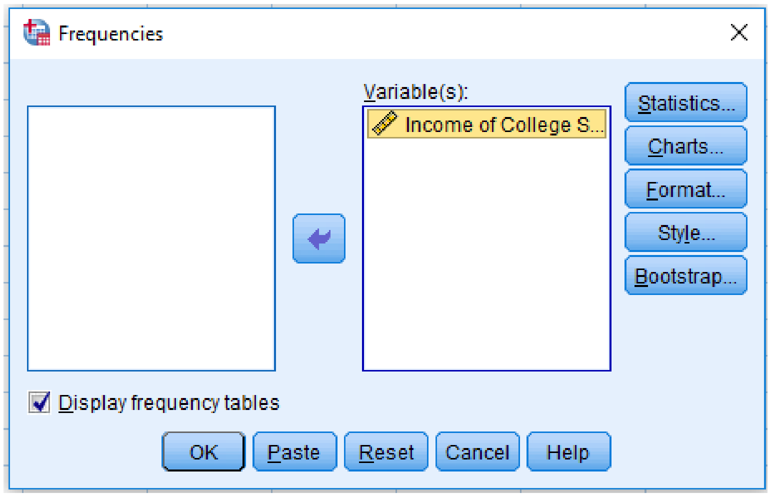


Income of College S...

☑ Display frequency tables

Step B.13 Click on the arrow symbol pointing to the right to move the variable label, 'Income of College Students', to the rectangle with the heading 'Variable(s)' (Figure 5.20).

Figure 5.20 Using the Frequencies Window

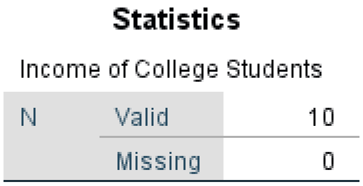


Variable(s): Income of College S...

☑ Display frequency tables

Step B.14 Click on '**OK**'. The result is the output shown in Figure 5.21, which includes the frequencies previously shown in Table 3.7 (the order is reversed) as well as additional information that may be of interest. This output can be printed and/or saved, if desired.

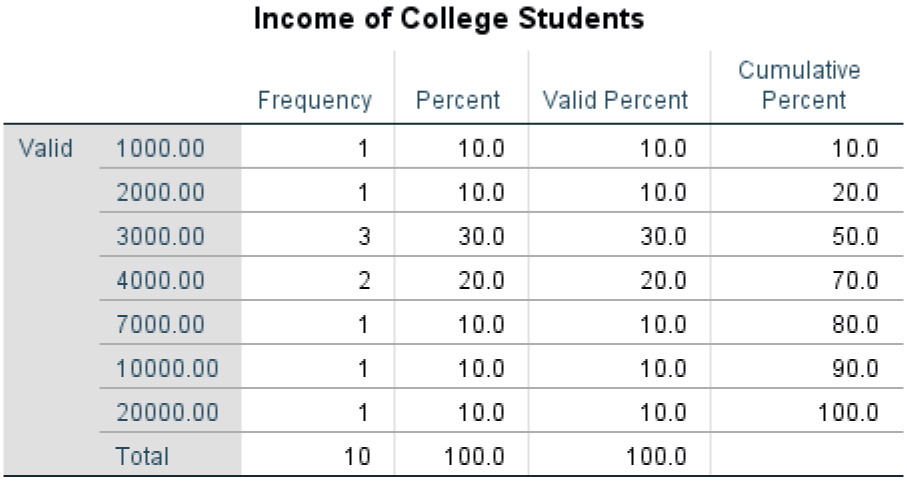
Figure 5.21 An Example of SPSS Output



Statistics

Income of College Students

N	Valid	10
	Missing	0



Income of College Students

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1000.00	1	10.0	10.0	10.0
	2000.00	1	10.0	10.0	20.0
	3000.00	3	30.0	30.0	50.0
	4000.00	2	20.0	20.0	70.0
	7000.00	1	10.0	10.0	80.0
	10000.00	1	10.0	10.0	90.0
	20000.00	1	10.0	10.0	100.0
	Total	10	100.0	100.0	

Step B.15 Exit from this output (not from SPSS). You will be prompted whether you want to save the output. As this is just an exercise, you do not have to save this output. You will have been returned to the window shown in Figure 5.17.

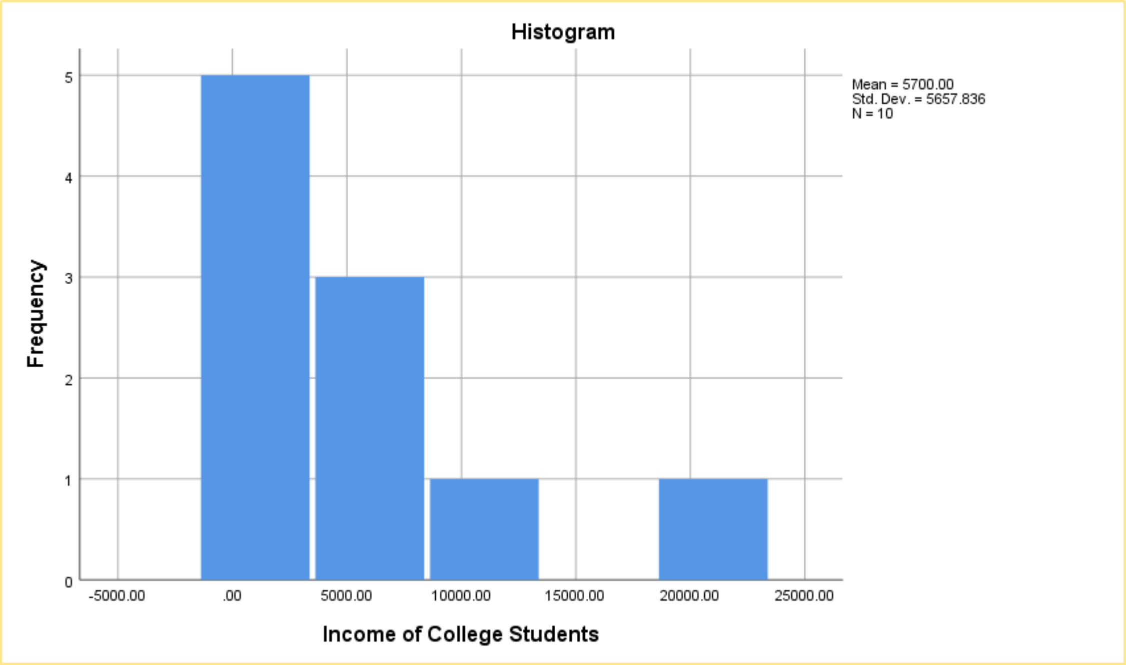
To Create A Histogram

Redo Steps B.10, B.11, B.12 and B.13.

Step B.16 Click on '**Charts**' and then '**Histograms**' and then '**Continue**'. If you do not want to see the frequency distribution again, click on the check in front of the phrase '**Display frequency tables**'.

Step B.17 Click '**OK**' and just the histogram in Figure 5.22 will appear. This is a somewhat condensed version of the histogram in Figure 3.1.

Figure 5.22 The SPSS Histogram



Step B.18 Exit from this output by clicking on the X in the upper right corner. You will be prompted whether you want to save the output. Once again, as this is just an exercise, you do not have to save this chart. If you continue exiting you will return to the Data View window. Exiting from this window will result in another prompt asking if you want to save the data. As this is just an example there is no need to save the data unless you feel you might want to use these data in the future.

A Third Example (C) Using SPSS

To Begin SPSS Proceed As In The Previous Example:

Step C.1 This is the same as step A.1.

Step C.2 This is the same as step A.2.

Step C.3 This is the same as step A.3.

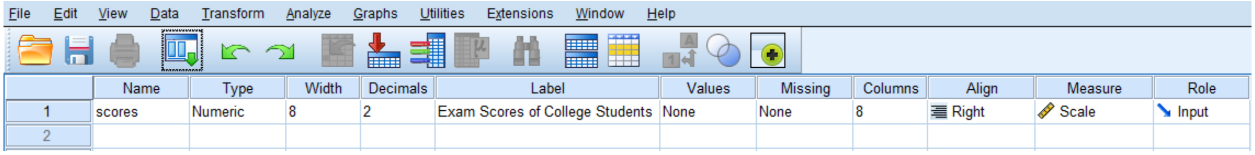
Step C.4 Click on the first empty cell under the column heading 'Name'. We are going to utilize the same data and labels as were previously employed in Table 3.10. As these data dealt with the exam scores of 10 hypothetical students, we have only one variable which I call 'scores'.

Step C.5 Click on the first empty 'cell' under the column heading 'Label'. In this cell I have typed 'Exam Score of College Students' as a more extensive description of our variable. Note that in order to see the entire label you will need to expand the size of this cell by placing your cursor on the right border of the Label heading and moving to the right.

Step C.6 In the current example all of the data are expressed as points on an exam. There is not, therefore, any need to label 'Values' of your variable.

Step C.7 Under the column heading 'Measure', choose 'Scale'. In SPSS, 'Scale' indicates that the data are at either the interval or ratio level of measurement. As our example deals with points on an exam, this is a ratio level of measurement and 'Scale' is the appropriate entry even though the data are discrete, not continuous. You have now completed the 'Variable View' window for the data we are interested in (Figure 5.23).

Figure 5.23 Defining a Label Within the Variable View Window

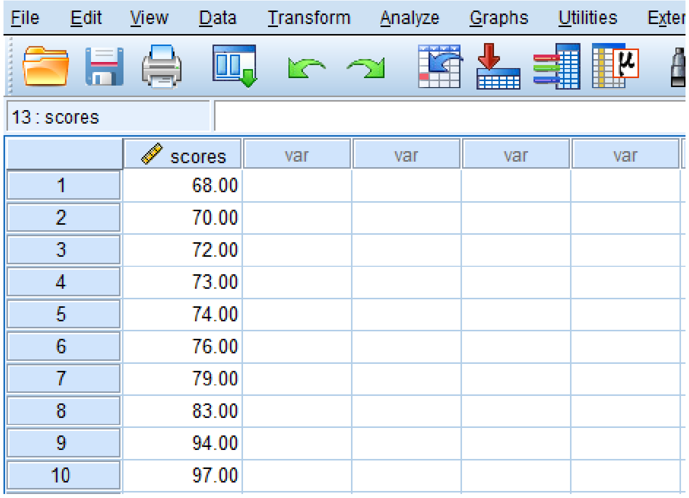


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	scores	Numeric	8	2	Exam Scores of College Students	None	None	8	Right	Scale	Input
2											

Step C.8 Click on the 'Data View' option at the lower left corner of the window. The variable 'scores' will now be visible.

Step C.9 Click on the first empty 'cell' under 'scores' and type in the number corresponding to the exam score of the first student, in this case 68. Continue adding data by clicking on the next empty 'cell' in the column under 'income' until all 10 values have been entered (Figure 5.24). You are now ready to use SPSS to describe the data you have entered.

Figure 5.24 Entering Data in the Data View Window



13 : scores

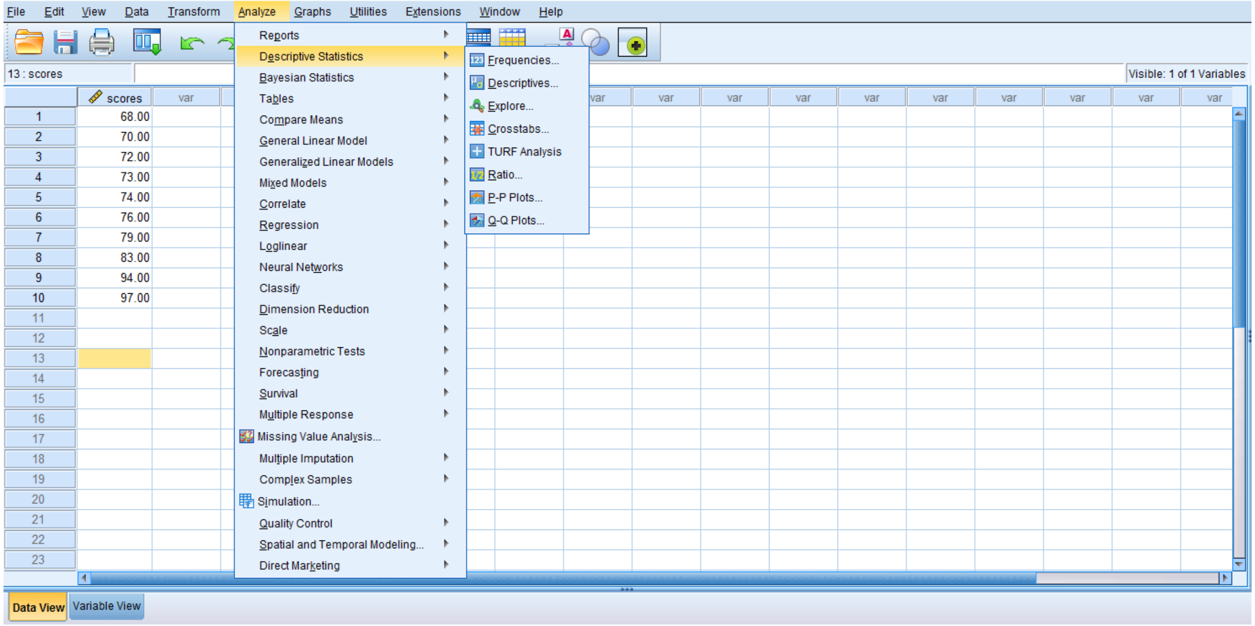
	scores	var	var	var	var
1	68.00				
2	70.00				
3	72.00				
4	73.00				
5	74.00				
6	76.00				
7	79.00				
8	83.00				
9	94.00				
10	97.00				

To Find Frequencies

Step C.10 Click on '**Analyze**' at the top of the Data View window.

Step C.11 Move your cursor down to '**Descriptive Statistics**'. When you do so, an additional window will appear (Figure 5.25).

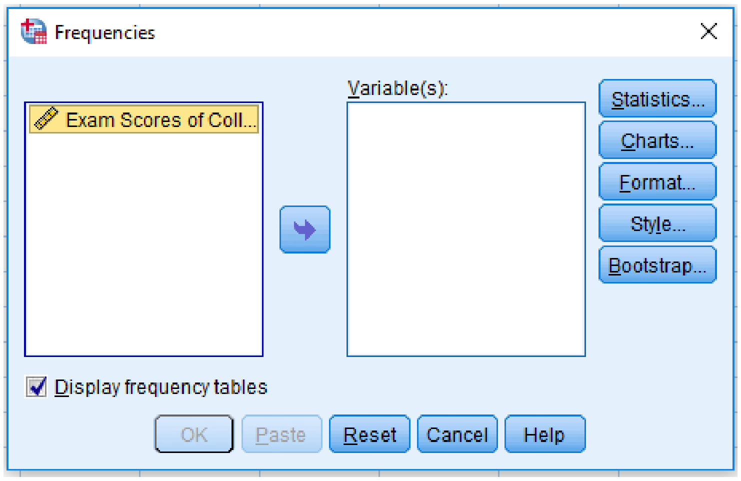
Figure 5.25 The Descriptive Statistics Function



	scores
1	68.00
2	70.00
3	72.00
4	73.00
5	74.00
6	76.00
7	79.00
8	83.00
9	94.00
10	97.00

Step C.12 Click on '**Frequencies**' and the window shown in Figure 5.26 will appear.

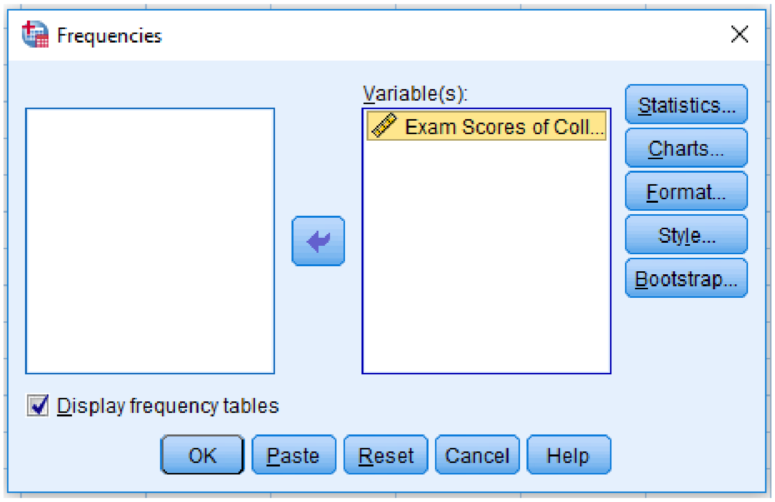
Figure 5.26 The Frequencies Window



☑ Display frequency tables

Step C.13 Click on the arrow symbol pointing to the right to move the variable label, 'Exam Scores of Coll...', to the rectangle with the heading 'Variable(s)' (Figure 5.27).

Figure 5.27 Using the Frequencies Window

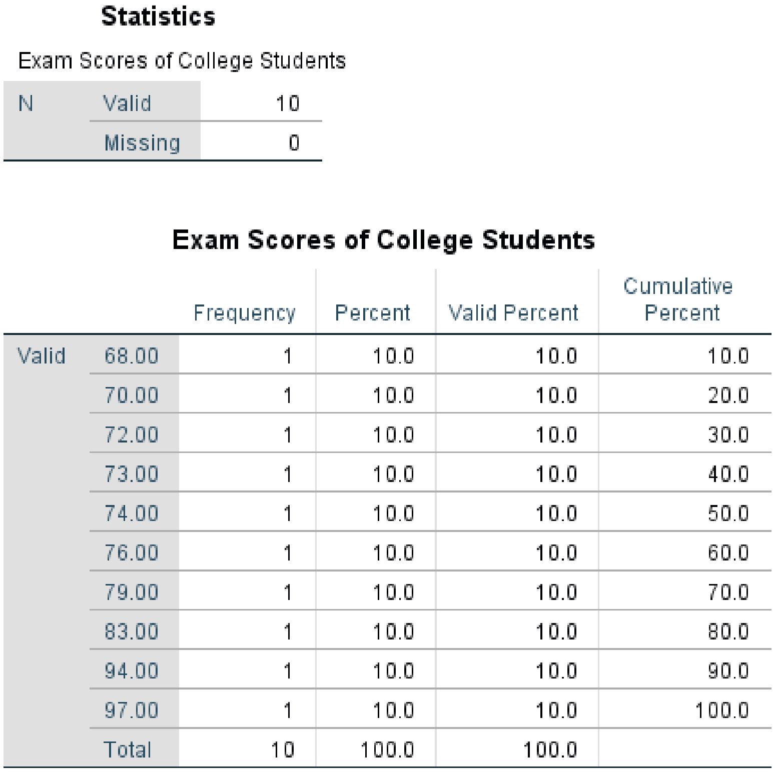


Variable(s): Exam Scores of Coll...

☑ Display frequency tables

Step C.14 Click on '**OK**'. The result is the output shown in Figure 5.28. This output can be printed if desired.

Figure 5.28 An Example of SPSS Output



Statistics

Exam Scores of College Students

N	Valid	10
	Missing	0

Exam Scores of College Students

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	68.00	1	10.0	10.0	10.0
	70.00	1	10.0	10.0	20.0
	72.00	1	10.0	10.0	30.0
	73.00	1	10.0	10.0	40.0
	74.00	1	10.0	10.0	50.0
	76.00	1	10.0	10.0	60.0
	79.00	1	10.0	10.0	70.0
	83.00	1	10.0	10.0	80.0
	94.00	1	10.0	10.0	90.0
	97.00	1	10.0	10.0	100.0
	Total	10	100.0	100.0	

Step C.15 Exit from this output (but not the SPSS program). You will be prompted whether you want to save the output. As this is just an exercise, you do not have to save this output. You will have returned to the window shown in Figure 5.24.

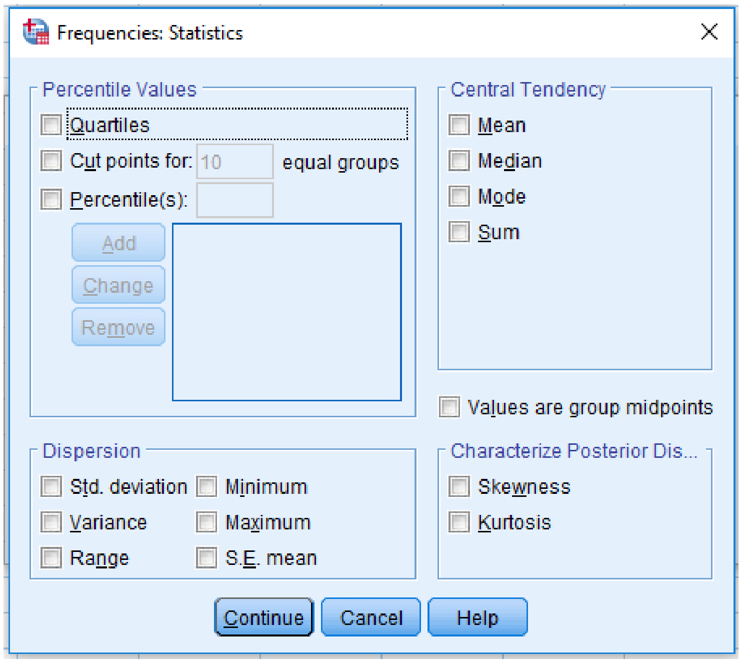
To Find Median And Quartiles

We will continue to use the same exam score data as previously.

Step C.16 Click on '**Analyze**', '**Descriptive Statistics**' and then '**Frequencies**'. The window shown in Figure 5.26 will appear.

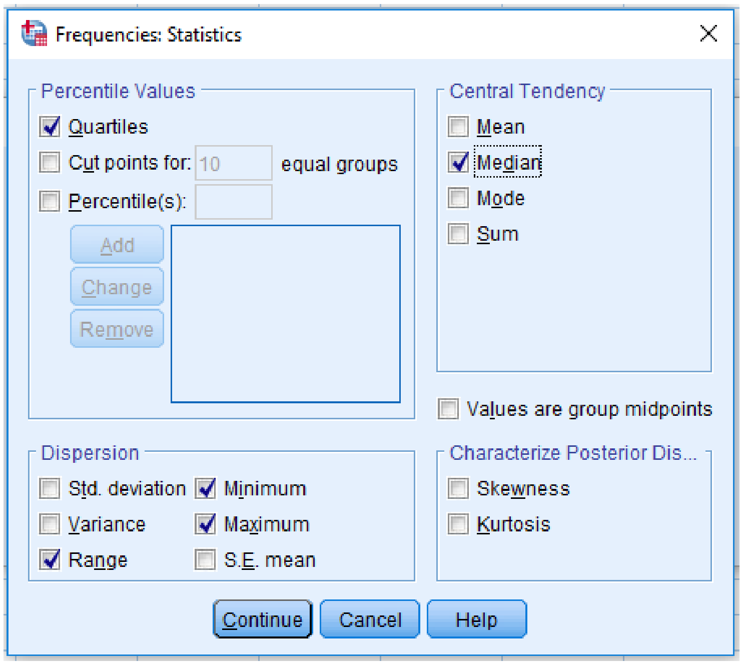
Step C.17 Click on the blue box '**Statistics**'. A new window will appear (Figure 5.29).

Figure 5.29 The Frequencies: Statistics Window



Step C.18 Now check the small boxes besides '**Quartiles**', '**Range**', '**Minimum**', '**Maximum**' and '**Median**'. The result will be Figure 5.30.

Figure 5.30 Calculation of the Quartiles and Median



Frequencies: Statistics

Percentile Values

☑ Quartiles

□ Cut points for: 10 equal groups

□ Percentile(s):

Central Tendency

□ Mean

☑ Median

□ Mode

□ Sum

□ Values are group midpoints

Dispersion

□ Std. deviation ☑ Minimum

□ Variance ☑ Maximum

☑ Range □ S.E. mean

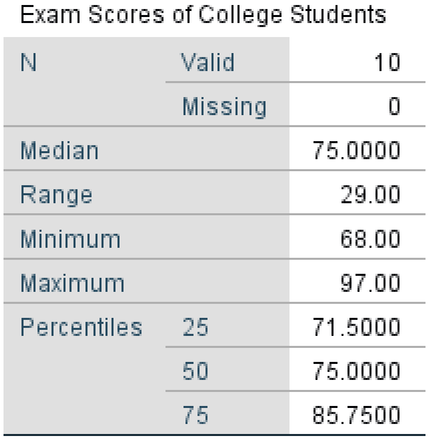
Characterize Posterior Dis...

□ Skewness

□ Kurtosis

Step C.19 Press '**Continue**'. This will return you to Figure 5.27. Now press '**OK**'. The output is shown in Figure 5.31.

Figure 5.31 Output of Calculating the Quartiles and Median



Exam Scores of College Students

N	Valid	10
	Missing	0
Median		75.0000
Range		29.00
Minimum		68.00
Maximum		97.00
Percentiles	25	71.5000
	50	75.0000
	75	85.7500

It is important to note that the range, minimum and maximum calculated by SPSS will be the same as we found previously in Chapter 3. However, *the 25th percentile (first quartile), 50th percentile (second quartile or median) and the 75th percentile (third quartile) may differ from what we obtained before.* This is because SPSS uses the original method proposed by Tukey for calculating these values while most of the field has gone to the simpler method that was described in Chapter 3. For problems in this text that you calculate by hand use the method described previously in Chapter 3. With problems solved using SPSS provide the values given by the computer package. Finally, if you are publishing the results of a study you have conducted, indicate that SPSS was used and then report the values from the program.

Step C.20 Exit from SPSS.

Conclusion

Hopefully you will agree that there isn't anything difficult about creating frequencies, bar graphs or pie charts with SPSS, or finding means, medians, modes, ranges, minimum and maximum values, or quartiles. The important points to remember are:

1. Each column in SPSS is a different variable.
2. Each row is a different subject.
3. Use the 'Variable View' window to name and label variables.
4. Use the 'Data View' window to enter data.
5. Use 'Analyze' to find descriptive statistics.
6. Use 'Charts' to create bar graphs, pie charts, and histograms. But remember, we use bar graphs with nominal or ordinal data, pie charts only with nominal data, and histograms with interval or ratio data.
7. Use 'Statistics' to calculate values for means, medians, modes, ranges, quartiles as well as minimum and maximum values. In the future you will learn that SPSS can also be used to find the standard deviation or variance of samples.

Glossary Of Terms

<u>Computational equations</u> – Equations developed to aid in statistical calculations. They were useful with large data sets, but now researchers would employ computer software packages instead.

<u>Data view</u> – SPSS window in which the data are displayed.

<u>SPSS</u> – A powerful, commonly-used statistical computer package. The letters 'SPSS' originally were an abbreviation for 'statistical package for the social sciences'.

<u>Variable view</u> – SPSS window in which variables are defined.

Questions – Chapter 5

(Answers are provided in Appendix J.)

1. In order to make calculation of large data sets easier, statisticians created _____.

a. a special, easily read type
b. computational equations
c. a unique form of slide rule
d. a rule to round off all decimals to whole numbers

2. If you were using SPSS and wanted to enter data to an already existing data file, you would go to the _____ window to enter a subject's responses.
a. Variable
b. Output
c. Data
d. Graphing

3. If you had already entered your data and now wanted to create a pie chart, which SPSS command would you begin with if you were at the data view window?
a. Analyze
b. Graphs
c. Compute
d. Pie Chart

4. Each column in the SPSS Data View window signifies a _____.
a. subject
b. experimental condition
c. different type of statistical analysis
d. variable

5. Each row in the SPSS Data View window signifies a _____.
a. subject
b. experimental condition
c. different type of statistical analysis
d. variable

6. In order to provide labels to clarify the meaning of the data, you would go to the _____.
a. Data View window
b. Variable View window
c. Analyze function
d. Graphs function

Problems 7 – 12 utilize SPSS

SPSS

With the following 13 scores, use SPSS to determine the range, as well as the first, second, and third quartiles:

1 2 3 3 5 6 7 9 10 12 16 22 60

7. The range is _____.
a. 13
b. 58
c. 59
d. 60

8. The first quartile is _____.
a. 3

b. 7
c. 14
d. 16

9. The second quartile is _____.
a. 3
b. 7
c. 14
d. 16

10. The third quartile is _____.
a. 3
b. 7
c. 14
d. 16

11. Twenty students take an exam in statistics and receive the following grades:

B	B	B	C
A	A	B	A
C	C	E	B
A	D	A	B
B	C	C	D

Enter these data in SPSS (using A = 4, B = 3, C = 2, D = 1 and E = 0) and then find the frequencies and make a bar graph.

12. Twenty five students report how many movies they have seen in the past week:

5	4	2	0	4
10	2	1	1	0
0	2	3	3	1
6	0	4	2	3
3	5	8	1	2

Enter your data in SPSS and then find the frequencies and make a histogram.

INFERENTIAL STATISTICS – MAKING DECISIONS BASED UPON YOUR DATA:

Chapter 6 – The Logic of Inferential Statistics

The Distinction between Difference and Association Questions

Chapter 7 – Finding Differences with Nominal Data – I

The Goodness-of-fit Chi-square

Chapter 8 – Finding Differences with Nominal Data – II

The Chi-square Test of Independence

Chapter 9 – Finding Differences with Interval and Ratio Data – I

The One-Sample z Test and the One-Sample t Test

Chapter 10 – Finding Differences with Interval and Ratio Data – II

The Independent Samples t and Dependent Samples t Tests

Chapter 11 – Finding Differences with Interval and Ratio Data – III

The One-way Between-subjects ANOVA

Chapter 12 – Finding Differences with Interval and Ratio Data – IV

The One-way Within-subjects ANOVA

Chapter 13 – Finding Differences with Interval and Ratio Data – V

The Two-way Between-subjects ANOVA

Chapter 14 – Identifying Associations with Interval and Ratio Data

The Pearson Correlation and Regression

Chapter 6
The Logic of Inferential Statistics: The Distinction between Difference and Association Questions

"By a small sample, we may judge of the whole piece."

Miquel de Cervantes, from Don Quixote

Where We Have Been

To this point in the book we have dealt with descriptive statistics, the procedures used to summarize data. More specifically, we have reviewed a number of procedures including how to construct frequency distributions, as well as how to calculate the measures that are employed for central tendency and variability. As you will recall, the frequency distribution provides an overview of the entire set of data, while the measures of central tendency and variability are single numbers that best summarize the data set's location and spread, respectively. As Table 6.1 indicates, the specific procedure chosen depends upon whether you are dealing with nominal, ordinal, or interval/ratio data. With the procedures that have been reviewed you are now in an excellent position to communicate a maximum amount of information in an efficient manner.

Table 6.1 An Overview of Descriptive Statistics (Summarizing Data)

	Type of Data			
	Nominal (Frequency)	Ordinal (Ranked)	Interval/Ratio (Score)	
Frequency Dist	Bar Graph or Pie Chart	Bar Graph	Histogram or Frequency Polygon	
Central Tendency	Mode	Median	IF NOT NORMAL Median	IF NORMAL Mean (Median – less common)
Variability	– – – –	Range	Interquartile Range	Standard Deviation z Score

Summary Presentation	Stem-and-leaf display and Boxplot	Stem-and-leaf display and Boxplot

Where We Are Going

The remainder of the text deals with inferential statistics. These procedures are used when making data-based decisions. Specifically, we will be concerned with whether the evidence for a relationship or pattern that is observed in a sample is sufficient to warrant concluding that the same relationship also exists in a population. In other words, we will be dealing with the question of whether a finding is likely to generalize. By the end of the book it will be evident that inferential statistical procedures are very powerful tools that can be used in a wide variety of situations.

Before beginning our discussion of inferential statistics it may be helpful to briefly review the distinction between a sample and a population. Recall that a population consists of all of the individuals that are potentially of interest. For instance, if we were interested in the effectiveness of a cancer therapy, the population would be all of the individuals who have cancer. Clearly, it is impractical to conduct a study that would examine every member of such a large population. Instead, we commonly select a sub-set of a population to examine. This sub-set is called a sample.

The sample should be chosen carefully. The goal is to select the members of the sample in such a way that any observed relationship in the sample can be generalized to the population that is of interest. For instance, let us assume that we are interested in what effect reducing legroom will have on passenger satisfaction on intercontinental airline flights. If the sample consists only of professional basketball players then it is questionable whether the findings will generalize to the population of all passengers. Usually, the optimal procedure for choosing a representative sample is to randomly select the subjects. In a **random sample** every member of the population has an equal chance of being chosen to be in the sample.

Random sample – A sample in which every member of the population has an equal chance of being chosen.

Decisions, Decisions, Decisions

For the remainder of the book you will be using samples to make decisions concerning populations. More specifically, you will be learning a set of agreed-upon procedures for making these decisions. Which procedure is appropriate depends upon the type of research design employed, as well as the type of data collected. As Table 6.2 indicates there are, fundamentally, two types of research questions, and therefore two broad types of research designs. **Difference designs** *examine whether an observed difference between samples is likely to have been the result of chance or, instead, provides evidence of a real or systematic effect*. If the experiment is properly designed and conducted, findings will generalize to the populations. This means that conclusions can be drawn regarding the populations based on sample data. There are a large number of difference designs. In addition, some difference designs provide a measure of interaction. We will be discussing this concept later.

Difference design – A research procedure designed to determine whether a difference observed between samples is likely to generalize to the populations.

Association designs are the other major type of research design. They *examine whether an association observed in a sample is likely to have been the result of chance or, instead, provides evidence of a systematic effect*. Once again, if the research is properly designed and conducted the findings based upon a sample will generalize to the corresponding population.

Association design – A research procedure designed to determine whether an association observed in a sample is likely to generalize to the population.

As Table 6.2 indicates, the choice of the appropriate statistical procedure is not only dependent on the research design, it is also a function of the measurement scale that was utilized. In other words, for inferential statistics, as for descriptive statistics, whether the data are nominal, ordinal, or interval/ratio is important in choosing the appropriate statistical procedure.

Finally, in Chapter 8 you will learn that the distinction between difference and association designs is not as clear when utilizing nominal data. Consequently, the same statistical procedure, the chi-square test of independence, can be utilized for either design (Table 6.2).

Table 6.2 An Overview of Inferential Statistics

	Type of Data			
	Nominal (Frequency)		Ordinal (Ranked)	Interval/Ratio (Continuous Measure)

When the Focus is on the Statistical Significance of a Difference:

Research Design		Research Design		
One Variable With At Least Two Outcomes	Goodness-of-fit Chi-Square	One IV With One Sample		One-sample z Test or One-sample t Test

		One IV With Two Or More Independent Samples	*Kruskal–Wallis H*[a]	One-way Between–Subjects ANOVA (Only two independent samples, Independent Samples t Test)
		One IV With One Sample Having Two Or More Repeated Measures		One-way Within–Subjects ANOVA (Only two repeated measures, Dependent Samples t Test)
Two Variables, Each With At Least Two Outcomes	Chi-Square Test of Independence	Two IV Each With Two Or More Independent Samples		Two-way Between–Subjects ANOVA

When the focus is on Characteristics and Statistical Significance of an Association:

Research Question

Association:	Chi-Square Test of Independence			
Correlation:	*Phi r*[b]		*Spearman r*[c]	Pearson r *Multiple Correlation*[d]
Regression:				Regression *Multiple Regression*[d]

Italicized items are reviewed in the following appendixes:

a. Appendix A
b. Appendix B
c. Appendix C
d. Appendix D

You may feel intimidated by Table 6.2. Don't be. The chapters that follow will walk you through each procedure. You mastered descriptive statistics and you will master inferential statistics. The goals of the remainder of this book are to assist you in understanding when it is appropriate to use each procedure, to illustrate how to conduct each of them, and to show how they are related.

A Little History

To appreciate the role of inferential statistics it is important to recognize that these procedures are a critical component of a process that has culminated in what we call the scientific method. Before the scientific method was developed there were other approaches for making decisions. Probably the most common is tradition, you do what has been done in the past. A major advance occurred with what is sometimes called **intrinsic plausibility**. In this system, alternative explanations are identified and then the explanation that is judged to be the most plausible or logical is accepted as being true. In many cases this approach is effective. Unfortunately, it is also likely that incorrect decisions will be made using this method. One of the major problems is that this method is susceptible to bias. At one time it was 'known' that the earth was the center of the universe and that the sun went around it. This was the most plausible explanation for the observation that the sun rose in the east every morning and set in the west every evening. Unfortunately, though plausible, it was incorrect. In addition, views of racial and sexual superiority and inferiority were, in the not too distant past, almost universally accepted. Those views seemed plausible to most people at the time. We no longer accept those views as logically defensible, and they are not supported by the relevant data. Numerous other examples could be cited.

Intrinsic plausibility – Decision-making process in which the alternative that seems most reasonable is accepted as being true.

By the end of the seventeenth century, a different approach to finding truth was evident in Western Europe. It is what we now call the **scientific method**. It has been refined into such a powerful approach that it may well be the most significant western contribution to world civilization. The scientific method seeks to identify relationships and express them mathematically. It relies upon a foundation of rigorous logic, but careful observation is the ultimate authority for determining truth. In other words, no matter how elegant the idea, if observations of nature do not provide support for it, the idea is rejected.

Scientific method – An approach to understanding that emphasizes rigorous logic, but also that careful observation is the ultimate authority for determining truth. It is a self-correcting approach that limits bias.

The scientific method does not ensure that every conclusion will be correct. However, a valuable feature of the scientific method is that it is self-correcting. What this means is that even if a conclusion is incorrect, or only partially correct, the process of scientific inquiry will, in time, provide a more adequate explanation. An example of this is the revolution that occurred in physics early in the 20th century. Newton's laws of motion had been proposed in the 17th century. They adequately accounted for everyday experience. However, the behavior of events at very high speeds, such as the speed of light, required a different explanation. This was provided by Einstein. Einstein's theory is able to account for more than Newton's. However, neither Newton's nor Einstein's theory can account for the behavior of very small, sub-atomic particles. As a result a great deal of effort is currently being devoted to developing an even more general theory.

Along with being self-corrective, the scientific method is efficient. One of the greatest advantages of this approach is that a researcher does not have to begin each inquiry from scratch. Instead, new research builds upon what has already been discovered. Research directions that are not productive are not pursued. Thus, effort is focused upon areas with the greatest potential for success. It is this economy of effort that has led to our increasingly rapid technological advances, increased life spans, higher standards of living and, unfortunately, the threat of nuclear war and global warming.

How Science Works

A critical component of the scientific method is making a comparison. Comparisons often involve two groups, or one group versus some standard. For instance, if we wanted to know if a coin was fair we would compare the results of tossing it a number of times to the expected outcome of an equal number of heads and tails. In this case, there would be only one set of data and we would compare it against what the data would be expected to be if the coin was fair.

It is important to note that some difference from the expected 50:50 ratio of heads and tails might have occurred even with a fair coin. After all, by chance we would expect to see some deviation from the predicted 50:50 ratio. So, how can we determine whether a difference that we observe in this ratio is the result of having an unfair coin, or is simply due to chance variation? Statisticians focus upon how great this difference is. Differences in ratios of heads and tails due to chance alone would be expected to be small. On the other hand, it is at least possible that an unfair coin would lead to a ratio of heads and tails that is substantially different from the expected 50:50 ratio.

Alternatively, let's assume we were interested in learning whether listening to a motivational speech would affect listeners' scores on an exam. We could begin by using random selection to obtain two samples of subjects which are likely to be equivalent. Then we could have the subjects in only one of the samples listen to the motivational speech. Finally, we would determine the scores on the exam for all of the subjects and check to see if there was now a difference between the two groups. In this between-groups study there would be two sets of data that would be compared. This experimental design can be summarized as the following:

Group 1 – no treatment
Group 2 – treatment

The group of subjects that *does not receive* the treatment is called the **control group**. The group of subjects that *does receive* the treatment is called the **experimental group**. While some difference between the two groups would be expected by chance, a large difference in the exam scores of the two groups would be evidence that the treatment had an effect.

<u>Control group</u> – In a between-groups design, the group of subjects that <u>does not receive</u> the treatment.

<u>Experimental group</u> – In a between-groups design, the group of subjects that <u>does receive</u> the treatment.

In this example, the researcher is in control of who listens to the motivational speech. The researcher then examines the subsequent exam scores. Thus two things are varying in this simple experiment. One variable, the presence or absence of the motivational speech, is determined or **manipulated** by the experimenter and is called the **independent variable** (IV). The experimenter wishes to make a decision whether the magnitude of the other variable, in this case the exam scores, is dependent upon the value of the IV that the subject receives. This second variable, which is *not* directly controlled by the experimenter, is called the **dependent variable** (DV).

Manipulate – The researcher determines which condition of the independent variable each subject receives.

Independent variable (IV) – In an experiment, the variable the experimenter manipulates or directly controls.

Dependent variable (DV) – In an experiment, the variable whose value is not directly controlled by the researcher. Its value may be changed by the independent variable (IV).

Making An Experimentally-Based Decision

"The only relevant test of the validity of a hypothesis is comparison of its predictions with experience."

Milton Friedman

We just noted that a treatment might, or might not have an effect. When designing a study we describe these possible outcomes by defining **hypotheses**. More specifically, with a difference design, the **null hypothesis**, symbolized as H_0, is usually that the treatment *does not* have an effect (the motivational speech does not have an effect on the exam scores). The **alternative hypothesis**, symbolized as H_1, is that the treatment *does* have an effect (the motivational speech does change the exam scores). The goal of an experiment is to permit the researcher to choose between these two mutually exclusive alternatives.

Hypothesis – A scientifically-based statement about some condition in the environment or population.

Null hypothesis (H_0) – When used with a difference design, the statement that the treatment does not have an effect.

Alternative hypothesis (H_1) – When used with a difference design, the statement that the treatment does have an effect.

To understand how a statistical procedure can assist in **hypothesis testing**, in other words in determining whether the null or alternative hypothesis is supported, let us return to our example of the experiment examining the effect of a motivational speech on exam scores. As explained previously, we would not be surprised if the mean exam scores of the control and experimental groups differed slightly. The question remains, how discrepant do the two means have to be in order to reject the null hypothesis that any difference is due to chance variation? In other words, how large a difference must be observed for the experimenter to conclude that the treatment had an effect on the exam scores? There is no absolute answer to this question. Statisticians provide an answer based upon the likelihood or probability of the null hypothesis being true. In most fields it has come to be accepted that if an outcome would be expected to occur, by chance, less than 1 time in 20 if the null hypothesis were true, then we reject the null hypothesis and accept the alternative hypothesis. One time in 20 is equivalent to .05 or 5%. There is nothing magical about .05. A different criterion such as .01 can be, and sometimes is, chosen. A criterion of .01 is equivalent to an outcome occurring only 1 time in 100 by chance. If a criterion of .01 is chosen, then we retain the null hypothesis unless an outcome is so unlikely that it would be expected to occur in less than 1 out of 100 cases by chance. We will discuss the reasons for using different criteria shortly. For now, it is important to understand that choosing a probability to differentiate between the null and alternative hypotheses is a critical part of experimentation. In fact, it is so important that the value that is chosen is given a name, the **alpha level** or **significance level**. Its symbol is the Greek letter alpha, α. As was just pointed out, the alpha level is commonly set at .05, but it could be some other value, such as .01.

Hypothesis testing – Statistically analyzing data to evaluate whether the null hypothesis should be retained or rejected.

Alpha level – Criterion set for rejecting the null hypothesis. This is usually .05.

Significance level – Another term for alpha level, the criterion set for rejecting the null hypothesis. This is usually .05.

In the example just given, the null hypothesis was that the treatment would not have an effect. The alternative hypothesis was that the treatment would have an effect. The null hypothesis would be rejected, therefore, if the mean for the experimental group's exam scores differs substantially from the mean of the control group. If we reject the null hypothesis, the alternative hypothesis will be accepted. However, if the mean of the experimental group is not so different from the mean of the control group, then the null hypothesis will be retained.

Whew, That Was A Lot To Remember

Summarizing to this point, when conducting an experiment we begin by defining a null hypothesis and an alternative hypothesis. These are mutually exclusive views of what the true situation is. We tentatively accept that the null hypothesis is true unless there is sufficient evidence from the experiment to indicate that this is unlikely. The criterion for deciding how unlikely the outcome must be in order to reject the null hypothesis is set by the experimenter when the alpha level is chosen. If the null hypothesis is rejected, we then tentatively accept that the alternative hypothesis is correct. We use the word 'tentatively' because it is important to recognize that with statistics we do not 'prove' that a difference that is observed between our samples will also exist between the corresponding populations. It is possible, in making an inference, that we have made an error. However, use of proper experimental and statistical procedures will minimize the likelihood of this happening.

Progress Check

1. In an experiment, the variable that the experimenter directly controls or manipulates is called the _____.
2. The possible outcome of an experiment that indicates that the treatment *does* have an effect is called the _____.
3. The criterion the experimenter sets for rejecting the null hypothesis is called the _____, and it is usually set at 1 chance in 20, or .05.

Answers: 1. independent variable 2. alternative hypothesis 3. alpha level

Probability, Error, And Power

"Absolute certainty is a privilege of uneducated minds – and fanatics.

It is, for scientific folk, an unattainable ideal."

Cassius J. Keyser

The decision-making process that we have been reviewing is based on the probabilities of outcomes. If an outcome is unlikely to have happened by chance we reject the null and accept the alternative hypothesis. Remember, by rejecting the null hypothesis we have not proven that it is incorrect. We are simply stating that, assuming alpha was set to .05, the odds are less than 5% that the observed difference happened by chance. It is possible, therefore, that we could be making an error when we reject the null hypothesis. In fact, we know the probability of rejecting the null hypothesis when it is actually true. That probability is alpha, which is usually 5%. In other words, if alpha is set at .05, then in 1 comparison out of 20 we will mistakenly reject the null hypothesis when it is in fact true. This is called **Type I error**.

<u>Type I error</u> – The probability of rejecting the null hypothesis when it is in fact true. This probability is equal to alpha, α, which is usually 5%.

No one likes to make errors. It might occur to you that it would be possible to reduce the probability of making a Type I error by simply reducing the size of alpha from .05 to .01. It is true that this step would reduce the Type I error rate. Unfortunately it would also have the unintended effect of increasing the probability of another type of error. By decreasing the Type I error rate, such as by setting alpha to .01 instead of .05, you make it harder to reject the null hypothesis. This is good if the null hypothesis is true. However, by decreasing alpha to .01 you simultaneously increase the likelihood that you will fail to reject the null hypothesis when it is in fact false. Failing to reject the null hypothesis when it is false is known as **Type II error**. Thus, as the probability of making a Type I error decreases, the probability of making a Type II error increases. The choice of the alpha level, therefore, is a compromise between these two types of error. And this choice is affected by which of these errors is felt to be most critical. Clearly, the consequences of failing to detect that a nuclear plant is unsafe are quite different than failing to detect that a new method of teaching chess is effective.

The probability of making a Type II error is called beta, and its symbol is the Greek letter β. The exact value of beta is usually not known. However, as was just discussed, what is known is that assuming nothing else in the experiment changes, if you reduce Type I error you will simultaneously increase Type II error. The reverse is also true.

<u>Type II error</u> – The probability of retaining the null hypothesis when it is in fact false. This probability is equal to beta, β. The probability of β is usually not known.

The relationship between Type I and Type II errors is shown in the top portion of Table 6.3.

Table 6.3 Relationship Between Type I and Type II Errors, and Power

		Experimenter's Decision	
		Rejects the Null Hypothesis	Retains the Null Hypothesis
Truth of the Decision (Which is not known.)	If Decision is Incorrect	*Type I error (α)* *Incorrectly rejected a true null*	*Type II error (β)* *Incorrectly retained a false null*
	If Decision is Correct	*Power (1 – β)* *Correctly rejected a false null*	*Correctly retained a true null*

Box dealing with Similarities Between Hypothesis Testing and Jury Decision Making

The material presented thus far in the chapter has been quite theoretical. You may find it helpful to realize that hypothesis testing is logically very similar to the decision-making process employed by juries (Feinberg, 1971). With each we have a system designed to lead to an informed conclusion. And each, broadly speaking, follows the same steps. As Table 6.4 indicates, we start with an initial assumption, then set a criterion for making our decision, have a selection process, define a basis for our decision, actually make a decision and, finally, we realize that we could have made an error. Concerning the errors, you should recognize that if a jury convicts an innocent defendant, then the jury has incorrectly rejected their initial assumption which was that the defendant was innocent. In hypothesis testing terms this would be an example of making a Type I error. Alternatively, if a jury finds a defendant not guilty when in fact the defendant committed the crime, then the initial assumption would have been incorrectly retained. In hypothesis testing terms this would be an example a making a Type II error.

You will see that with hypothesis testing we have additional terms to learn, and they will require attention in order to be mastered. But hopefully you recognize that the logic of hypothesis testing has parallels to a process that you may already be familiar with from watching TV and reading the news. Just be cautious not to overstate the similarities (Martin, 2003).

Table 6.4 Comparison of Jury Decision Making with Hypothesis Testing

	Step 1 Initial Assumption	Step 2 Criterion for Decision	Step 3 Selection Process	Step 4 Basis for Making Decision	Step 5 Decision	Step 6 Not a Proof
Jury Decision Making	defendant assumed to be innocent	beyond reasonable doubt	impartial jury chosen	testimony in court	defendant found innocent or guilty	incorrect decision is possible
Hypothesis Testing	null hypothesis assumed to be true	alpha level	randomly selected samples	data from samples	accept or reject null hypothesis	Type I or Type II error is possible

Since errors can occur with scientific hypothesis testing we are faced with a dilemma. We want to make the correct decision, of course, but if we reduce one type of error, we simultaneously increase the probability of the other type. Is there anything that we can do to increase the likelihood of coming to a correct decision? When statisticians address this issue, they introduce the concept of **power**. Power is simply the probability of correctly rejecting a false null hypothesis. Fundamentally, *the goal of experimentation is to conduct as powerful a study as possible*.

The probability of power is 1 – β. Of course, we just learned that we rarely know the precise probability of β, so we usually do not know the probability of 1 – β either. However, the concept of power is still very useful. How power is related to an experimenter's research decision is illustrated in Table 6.3. (The fourth cell in the table, the probability of correctly retaining a true null hypothesis is not of interest to us at this point). It is important to recognize that even though we usually do not know the probability of β, we can nevertheless take steps to increase the probability of correctly rejecting a false null hypothesis, in other words to increase the power of our study. Some of these steps are listed in Table 6.5.

<u>Power</u> – The probability of correctly rejecting a false null hypothesis. The probability is 1 – β.

Table 6.5 Some Steps That Will Increase Power

1. Pick a treatment that is likely to have a large effect.
2. Choose a measurement scale that has as much information as possible.
3. Increase the alpha level (e.g., from .01 to .05).
4. Increase the sample size.
5. Conduct the study so that sources of unwanted variability are minimized.

As Table 6.5 indicates, one step that the researcher can take to increase the power of an experiment is to choose a treatment level that is likely to cause a noticeable effect. In some cases this can be determined from previous, successful studies. In other cases the researcher will have to make an educated guess. For instance, if you are studying the effect of sleep deprivation on the ability to do math calculations, it is not likely that you will find an effect if the sleep deprivation consisted of only a loss of 15 minutes of sleep. You would be more likely to find an effect if the size of this intervention were much greater. Perhaps a deprivation of 2 or 3 hours would be a better choice.

A second step that you may be able to take is to choose a measure that uses interval or ratio data. The statistical tests that are employed with interval or ratio data are more efficient than those that use nominal or ordinal data. This means that you will not need as many subjects to detect the same size effect if you use interval or ratio data. If you cannot use interval or ratio data, your next best choice would be to use ordinal rather than nominal data.

Choosing alpha to be .05 rather than .01 will also increase the likelihood of rejecting the null hypothesis when in fact it is false. Of course, by taking this action you will increase the probability of making a Type I error, but 5% is usually an acceptable level for this type of error in the social sciences.

All else being equal, having larger samples will make rejecting the null hypothesis easier. For instance, with interval or ratio data the means of larger samples are expected to vary less from each other, by chance, than the means of smaller samples. As you will see later in the book, a consequence is that with large samples you do not need as large a difference between the experimental and control groups in order to reject the null hypothesis.

Finally, any steps that you can take to reduce unwanted variability while conducting the study will increase the power of your study. Among the steps to be considered are using consistent procedures with all of the subjects and controlling conditions during testing, such as the temperature and humidity, that might affect the outcome.

Real-World Limitations

As the previous overview of the scientific method indicated, the optimal features of an experiment include random assignment of subjects and manipulation of the independent variable by the researcher. When both random assignment and manipulation of the independent variable have occurred we have what is called a **true experiment**. And assuming the research has been carefully conducted, the experimenter is justified in rejecting the null hypothesis and accepting the alternative hypothesis based upon the outcome of the study. More specifically, the researcher can come to what is called a **cause-and-effect conclusion**; the change in the value of the independent variable resulted in a change in the value of the dependent variable. Of course, as was noted previously, this is a probabilistic decision. It is always possible that a Type I error has occurred. However, the probability of making a Type I error, called alpha, is known and is small.

An example of a true experiment is the classic Bobo doll research begun by Bandura (1961). This research, which has involved a number of studies, examined imitation of aggression in children. In these studies children were randomly assigned to either the control group or one of several experimental groups. The experimenter controlled whether the child observed aggressive or non-aggressive behavior (there were a number of conditions). Then, the behavior of each child was recorded. It was found that when they were frustrated many of the children imitated the specific aggressive behavior they had previously witnessed. This is an example of a true experiment for there is random assignment of the subjects and experimenter manipulation of the independent variable.

True experiment – An experiment in which the researcher randomly assigns the subjects and also manipulates the value of the independent variable. As a result, at the conclusion of the study the researcher is justified in reaching a cause-and-effect conclusion concerning the relationship between the independent and dependent variables.

Cause-and-effect conclusion – Decision that the change in the value of the independent variable resulted in a change in the value of the dependent variable. This is justified with a well-conducted, true experiment.

However, in many real-world situations it is not possible for an experimenter to randomly assign subjects and manipulate the variable that is of interest. In these cases the scientific method can still be employed, but the strength of the conclusion that the researcher is justified in making is reduced. For instance, if the researcher can manipulate the independent variable but cannot randomly assign the subjects, then the study has some, but not all, of the characteristics of a true experiment. Accordingly, it is called a **quasi-experiment**. Compared to a true experiment we are now less confident that a difference found at the conclusion of the study is due to the manipulation of the independent variable. For example, in a classic series of studies Gazzaniga (1967) examined people whose corpus callosum (a band of neurons connecting the two hemispheres of the brain) had been surgically cut in order to prevent the spread of seizure activity. These subjects were asked to identify objects placed in either of their hands while they were prevented from looking. When the object was placed in the right hand, the subjects could name it. However, when the object was placed in the left hand, they could not name it. As the neural inputs to the brain cross this is evidence that the ability to verbally name objects is lateralized to the left hemisphere. You should recognize that this is a quasi-experimental design. The subjects were not being randomly assigned. Instead, they came to the study with or without having had their corpus callosum cut. However, the experimenter was in control of the independent variable, the placement of the objects to be named. The results of these studies were dramatic. And no one is going to seriously suggest that Gazzaniga should instead have conducted a true experiment, for this would have required random assignment of subjects to undergo surgery!

Quasi-experiment – An experiment in which some characteristic of a true experiment is missing. Most commonly, the researcher manipulates the value of the independent variable but does not randomly assign the subjects. As a result, at the conclusion of the study the researcher has less confidence in concluding that there is a cause-and-effect relationship between the independent and dependent variables than would be the case with a true experiment.

Alternatively, it may not be possible for the researcher to either randomly assign the subjects or to manipulate a variable. This is called a **correlational study**. Due to the lack of control we have even less confidence concerning the cause of any obtained relationship. As the researcher did not manipulate any variable there is not an independent variable, or a dependent variable. Either variable could be causing a change in the other, both could be affecting each other, or some other variable(s) could be affecting them both. For example, it is commonly noted that football teams that have a propensity to turn the ball over to the other team are also more likely to lose the game. This may seem to be an obvious cause-and-effect relationship; repeatedly giving the ball to the opponent causes an increase in the likelihood of losing the game. However, it is important to recognize that this is a correlation, for there is neither random assignment nor experimenter control of a variable. And upon closer inspection the interpretation of the relationship becomes somewhat less certain. For instance, the likelihood of turnovers increases if a team is passing rather than running when they have possession of the ball. And teams that are already far behind are more likely to rely upon passing the ball as a desperate means to score quickly. In other words, what seemed initially like an obvious cause-and-effect relationship, turnovers cause teams to lose, is more complex as teams that are already losing are also more likely to have turnovers!

Correlational study – A study in which the researcher does not randomly assign the subjects and does not manipulate the value of a variable. As a result, at the conclusion of the study the researcher has little confidence that there is a cause-and-effect relationship between the variables.

Conclusion

The first section of this book dealt with descriptive statistics. These are the procedures that we use to summarize a set of data. This chapter introduced inferential statistics. Inferential statistics are a set of procedures that assist us in determining whether a relationship or pattern observed with a sample(s) is likely to generalize to a population(s). The remainder of the book will be dealing with inferential statistical procedures.

No new statistical procedures were introduced in this chapter. Instead, the emphasis was on reviewing the logic of hypothesis testing. The remainder of the text will build upon this foundation. It is important, therefore, that you master the concepts and the associated terms described in this chapter. Specifically, the null and alternative hypotheses were defined in the context of a between-groups experiment. Then, the rationale for making probabilistic decisions, and thus the necessity of choosing an alpha level, was covered. It was noted that statistical procedures assist in the decision-making process but do not ensure that every conclusion will be correct. This led to a discussion of Type I and Type II errors. The steps that a researcher can take that will increase the power of a study were then briefly described. Finally, the distinctions between a true experiment, quasi-experiment, and a correlational study were reviewed.

It should be noted that the logic of hypothesis testing applies equally well to studies that employ nominal, ordinal or interval/ratio measurement scales. In the next chapter we will turn our attention to the inferential procedures that are used with nominal data. Subsequent chapters will describe the procedures utilized with interval/ratio data (procedures utilized with ordinal data are reviewed in the appendixes). We are, therefore, continuing with the same general organization that we employed with our discussion of descriptive statistics.

Glossary Of Terms

Alpha level – Criterion set for rejecting the null hypothesis. This is usually .05.

Alternative hypothesis (H_1) – When used with a difference design, the statement that the treatment does have an effect.

Association design – A research procedure designed to determine whether an association observed in a sample is likely to generalize to the population.

Cause-and-effect conclusion – Decision that the change in the value of the independent variable resulted in a change in the value of the dependent variable. This is justified with a well-conducted, true experiment.

Control group – In a between-groups design, the group of subjects that does not receive the treatment.

Correlational study – A study in which the researcher does not randomly assign the subjects and *does not manipulate the value of a variable. As a result, at the conclusion of the study the researcher has little confidence that there is a cause-and-effect relationship between the variables.*

Dependent variable (DV) – In an experiment, the variable whose value is not directly controlled by the researcher. Its value may be changed by the independent variable (IV).

Difference design – A research procedure designed to determine whether a difference observed between samples is likely to generalize to the populations.

Experimental group – In a between-groups design, the group of subjects that does receive the treatment.

Hypothesis – A scientifically-based statement about some condition in the environment or population.

Hypothesis testing – Statistically analyzing data to evaluate whether the null hypothesis should be retained or rejected.

Independent variable (IV) – In an experiment, the variable the experimenter manipulates or directly controls.

Intrinsic plausibility – Decision-making process in which the alternative that seems most reasonable is accepted as being true.

Manipulate – The researcher determines which condition of the independent variable each subject receives.

Null hypothesis (H_0) – When used with a difference design, the statement that the treatment does not have an effect.

Power – The probability of correctly rejecting a false null hypothesis. This probability is 1 – β.

Quasi-experiment – An experiment in which some characteristic of a true experiment is missing. Most commonly, the researcher manipulates the value of the independent variable but does not randomly assign the subjects. As a result, at the conclusion of the study the researcher has less confidence in concluding that there is a cause-and-effect relationship between the independent and dependent variables than would be the case with a true experiment. .

Random sample – A sample in which every member of the population has an equal chance of being chosen.

Scientific method – An approach to understanding that emphasizes rigorous logic, but also that careful observation is the ultimate authority for determining truth. It is a self-correcting approach that limits bias.

Significance level – Another term for alpha level, the criterion set for rejecting the null hypothesis. This is usually .05.

True experiment – An experiment in which the researcher randomly assigns the subjects and also manipulates the value of the independent variable. As a result, at the conclusion of the *study the researcher is justified in reaching a cause-and-effect conclusion concerning the relationship between the independent and dependent variables.*

Type I error – The probability of rejecting the null hypothesis when it is in fact true. This probability is equal to alpha, α, which is usually 5%.

Type II error – The probability of retaining the null hypothesis when it is in fact false. This probability is equal to beta, β. The probability of β is usually not known.

References

Bandura, A., Ross, D., & Ross, S. A. (1961). Transmission of aggression through imitation of aggressive models. *Journal of Abnormal and Social Psychology, 63,* 575-582.

Feinberg, W. E. (1971). Teaching type I and type II errors: The judicial process. *The American Statistician, 25*(3), 30-32.

Gazzaniga, M. S. (1967). The split brain in man. *Scientific American, 217*(2), 24-29.

Martin, M. A. (2003). "It's like … you know": The use of analogies and heuristics in teaching introductory statistical methods. *The Journal of Statistics Education, 11*(2) (www.amstat.org/publications/jse/v11n2/martin.html)

Questions – Chapter 6

(Answers are provided in Appendix J.)

1. In an experiment, the group that does *not* receive the treatment is called the _____ group.
   a. alpha
   b. benign
   c. control
   d. critical

2. The probability of correctly rejecting a false null hypothesis is called _____.
   a. alpha
   b. beta
   c. error rate
   d. power

3. The essential feature of random assignment is that _____.
   a. every member of a population has an equal probability of being chosen
   b. no one knows who will be chosen
   c. subjects are clueless as to the purpose of the experiment
   d. only volunteers take part in a study

4. In an experiment, the _____ is that the treatment does have an effect and the _____ is that it does not have an effect.

a. null hypothesis; alternative hypothesis
b. alternative hypothesis; null hypothesis

5. The experimenter sets the probability of _____ but usually does not know the probability of _____.
a. alpha; beta
b. Type II error; Type I error
c. beta; alpha
d. Type I error; alpha

6. If you reject the null hypothesis when in fact it is true, you have _____.
a. made a Type I error
b. broken the law and will need a lawyer
c. made a Type II error
d. shown that you have mastered experimental methodology

7. To increase power, an experimenter would _____.
a. decrease the sample size.
b. choose a nominal rather than an interval measurement scale.
c. increase the alpha level (e.g., from .01 to .05).
d. not use any of the above options.

8. The criterion for rejecting the null hypothesis is set by the experimenter and in the social sciences is usually equal to _____.
a. .10
b. .05
c. .01
d. .001

9. Critical features of the scientific method include all of the following *except* _____.
a. ultimately relies upon observation
b. requires careful, rational thought
c. never is in error
d. often makes use of experiments

10. The experimenter usually will *not* know which of the following?
a. significance level
b. alpha level
c. value of beta
d. the number of subjects in the experiment

11. There are, fundamentally, two types of research questions, and therefore two types of research designs. These are called _____ and _____ designs.
a. error prone; truthful
b. small scale; large scale
c. plausible; implausible
d. difference; association

12. The ancient Greeks are famous for employing _____ to determine truth.
a. intrinsic plausibility
b. the scientific method
c. statistical analysis

d. difference designs

13. The probability of making a Type II error is _____.
    a. alpha
    b. the criterion set by the experimenter
    c. equal to the region of rejection
    d. beta

14. A basic goal of experimentation is to conduct _____.
    a. as complex a study as possible
    b. as powerful a study as possible
    c. a study with a criterion of .10.
    d. a study where Type I error has been eliminated

15. A professor is interested in improving her students' grades, and tests whether having the students engage in light exercise will be beneficial. In this example, the independent variable is _____ and the dependent variable is _____.
    a. students' grades; exercise
    b. exercise; students' grades

16. In a quasi-experiment the experimenter commonly manipulates the _____ but does not _____.
    a. dependent variable; manipulate the independent variable
    b. independent variable; randomly assign the subjects
    c. control group; manipulate the experimental group

17. In a correlational study the experimenter _____.
    a. does not manipulate the independent variable or randomly assign the subjects
    b. does manipulate the independent variable and does randomly assign the subjects
    c. never makes a Type I error
    d. never makes a Type II error

18. In a true experiment the researcher _____.
    a. manipulates the independent variable
    b. randomly assigns the subjects
    c. always rejects the null hypothesis
    d. both a and b, but not c

19. Assuming no other aspect of the experiment changes, if the probability of Type I error is decreased from .05 to .01, the probability of Type II error will _____.
    a. decrease
    b. stay the same
    c. increase

20. The strongest statement concerning the relationship of two variables can be made by a researcher following a _____.
    a. true experiment
    b. quasi-experiment
    c. correlational study
    d. all lead to statements of equivalent strength

Chapter 7
Finding Differences with Nominal Data – I: The Goodness-of-Fit Chi-Square

Statistics may be defined as "a body of methods for making
wise decisions in the face of uncertainty."
W. A. Wallis

Introduction

We begin our exploration of inferential statistical procedures with a focus upon the simplest level of measurement. Recall that nominal data are discrete and refer to categories. These data consist only of frequencies. An example of nominal data would be if you were to determine how many members of a group consider themselves to be Republicans, how many consider themselves to be Democrats, how many have some other party affiliation, and how many have no political affiliation at all. You would simply determine the frequency in each category.

In this and the next chapter we will be utilizing procedures that do not make assumptions about a population's parameters, such as its variability, and do not assume that the population is normally distributed. They are thus nonparametric as well as distribution-free, but we will follow convention and simply refer to them as **nonparametric procedures**. Later, when we are dealing with interval and ratio data, we will study tests that do make assumptions about population parameters and distributions. They are called **parametric procedures**. In this chapter we will begin our discussion of the nonparametric inferential procedures with the goodness-of-fit chi-square test. This test is underlined in Table 7.1.

Nonparametric procedure – Statistical procedure that does not make assumptions about the population's parameters and does not assume that the population is normally distributed.

Parametric procedure – Statistical procedure that does make assumptions about the population's parameters and does assume that the population is normally distributed.

Table 7.1 Overview Table of Inferential Statistical Procedures For Finding if there is a Difference

Type of Data		
Nominal	Ordinal	Interval/Ratio
(Frequency)	(Ranked)	(Continuous

				Measure)
Research Design		Research Design		
One Variable With At Least Two Outcomes	Goodness-of-fit Chi-Square	One IV With One Sample		One-sample z Test or One-sample t Test
		One IV With Two Or More Independent Samples	*Kruskal–Wallis H*	One-way Between–Subjects ANOVA (Only two independent samples, Independent Samples t Test)
		One IV With One Sample Having Two Or More Repeated Measures		One-way Within–Subjects ANOVA (Only two repeated measures, Dependent Samples t Test)
Two Variables, Each With At Least Two Outcomes	Chi-Square Test of Independence	Two IV Each With Two Or More Independent Samples		Two-way Between–Subjects ANOVA

The Italicized procedure is reviewed in Appendix A

Goodness-of-Fit Chi-Square

Assume that you have a coin and you want to determine whether it is 'fair'. Of course a fair coin should land heads 50 percent of the time and tails 50 percent of the time. However, some deviation from the expected 50:50 split would not be surprising. After all, if you tossed a coin a hundred times and found that there were 49 heads and 51 tails, most observers would say that this is close enough to the expected 50:50 proportion. But this raises the question of how far from 50:50 you would have to be in order to reject the view that the coin is fair and accept the alternative that the coin is biased. Hopefully you recognize that we have just stated a null and an alternative hypothesis. In Chapter 6, we learned that the null hypothesis is a statement of no effect. In our current case we assume that there will be no difference in the number of heads and tails for the population consisting of all tosses of the coin. The coin will thus land with a 50:50 split in a long series of tosses. The alternative hypothesis is that there is a difference in the number of heads and tails for the population consisting of all tosses of the coin. In this case there will be a difference in the number of heads and tails in our sample, and thus the proportion of heads to tails will deviate from the expected 50:50 ratio. The question for us is how great the difference must be in a sample in order for us to reject the null hypothesis that this difference is due to chance, and instead accept the alternative hypothesis that the discrepancy is indicative of the coin being biased. In science we want a process that will lead to a consistent decision regardless of who the decision-maker is. In other words, we cannot just leave it up to each individual to decide whether a particular conclusion is plausible.

This may all seem unnecessary, even unimportant, but it is not. Decisions matter and the process by which they are derived is critical. In everyday life we are used to being rather 'sloppy' in our decision making. We are all affected by our emotions and on occasion we make decisions based on too little information. As a result we make mistakes. Several approaches have been developed to reduce the likelihood of coming to erroneous decisions. In philosophy this includes the study of logic. In science there are a set of procedures known collectively as the scientific method. A critical component of the scientific process is that we base our decisions upon the evidence that we have collected. We then employ accepted statistical procedures to assist in arriving at a decision. Thus, an advantage of the scientific approach is that it assures an outcome that is more than just someone's opinion.

For example, let's assume that your kind, thoughtful professor meets you in the hall one day before class. As both of you have come to class early there is time to engage in stimulating intellectual conversation. Instead, and hopefully unrealistically, your professor suggests that you pass the time by wagering on the outcome of coin tosses, and he/she just happens to have a favorite coin to use. You, of course, cannot imagine that your professor would be anything less than scrupulously honest, so you accept the offer to bet your hard-earned lunch money. Your professor indicates that he/she is rather partial to heads. You do not mind. Why should you? After all, tails should come up as often as heads, assuming of course that the coin is fair.

At the end of 10 tosses, you note that while you have won 3 times, your professor has won 7 times. What should you think? While hopefully not a likely situation, the implications of your decision should be clear. If you retain the null hypothesis, which in this case would be that any discrepancy from the expected 50:50 outcome is due to chance, then you might continue to play the game and there would be no reason to accuse your professor of engaging in dishonest behavior. However, if you accept the alternative hypothesis that the obtained proportion of heads and tails differs from what would be expected if the coin was fair, then you might conclude that your professor has knowingly engaged in dishonest behavior. It would be very awkward for you to accuse your professor of dishonesty based simply upon your personal opinion. You would want to be on firmer footing just in case the department chair or the dean happened to get involved. The issue is straightforward. Quite simply, is a 7 to 3 outcome different enough from the expected 5 to 5 outcome to warrant the conclusion that your professor is using a biased coin?

To answer this question we must employ an inferential statistical procedure appropriate for finding if there is a difference. As the overview table (Table 7.1) indicates, there are a number of procedures to choose from.

Fortunately, the process for choosing the correct procedure is straightforward. We first note that we have nominal data and that there is only one variable of interest (the outcome of flipping the coin). Further, this variable has two possible outcomes, either heads or tails. Thus, referring to Table 7.1 we find that the **goodness-of-fit chi-square test** (which has the symbol χ^2) would be appropriate (this procedure is underlined in the table). Specifically, the goodness-of-fit chi-square test uses frequency data from one variable, in this case the outcome of flipping a single coin, to test whether the proportion that has been obtained differs from the proportion that would be expected if the null hypothesis were correct. With our example we expect there to be 5 heads and 5 tails in 10 tosses of a fair coin.

Goodness-of-fit chi-square test – An inferential procedure that tests whether observed frequencies differ from expected frequencies.

Step 1: State the null and alternative hypotheses, and specify the alpha level (In an experiment this step would occur before any data are collected.):

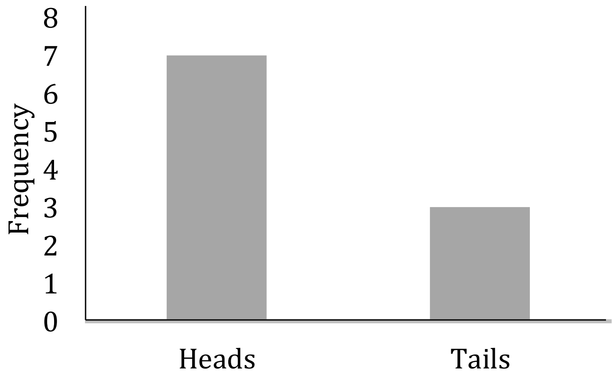
H_0 – The coin is not biased (it is fair); any observed deviation from the expected 50:50 outcome is due to chance.

H_1 – The coin is biased; any observed deviation from the expected 50:50 outcome is not due to chance.

Alpha is set to .05.

Our data can be summarized using a bar graph (Figure 7.1).

Figure 7.1 Example 1: Bar Graph of Coin Tosses



Inferential statistical procedures, such as the goodness-of-fit chi-square test, are based upon assumptions. One of the assumptions of the chi-square test is that no observed event influences another. This assumption is reasonable with our example for it is simply indicating that the test requires that the outcome of one flip of the coin does not affect the outcome of any other flip. In statistical terms, we would say that the outcomes are **independent**. (This concept will be discussed in more detail in Chapter 8.)

<u>*Independent*</u> *– Two events, samples or variables are independent if knowing the outcome of one does not enhance our prediction of the other.*

The goodness-of-fit chi-square test compares the **observed frequencies** with the **expected frequencies**. If the null hypothesis of no difference in the observed and expected frequencies is correct, then there should be a close correspondence between these two sets of frequencies in our data. Specifically, in our example of the coin toss we should get a proportion close to 50:50. Of course, some discrepancy is likely. The issue is whether our result of 7 heads and 3 tails is so unlikely to have happened by chance that we should reject the null and accept the alternative hypothesis that the coin is biased.

<u>*Observed frequencies*</u> *– With nominal data, the actual data that were collected.*

<u>*Expected frequencies*</u> *– With nominal data, the outcome that would be expected if the null hypothesis were true.*

The calculation of the goodness-of-fit chi-square is straightforward, though it may not appear to be at first glance:

$$\text{Chi-square} = \chi^2 = \Sigma\,\frac{(\text{Frequency observed} - \text{Frequency expected})^2}{\text{Frequency expected}} = \sum\frac{(\mathrm{f_o} - \mathrm{f_e})^2}{\mathrm{f_e}}$$

where f_o = frequency observed, and f_e = frequency expected.

Though this equation may look intimidating it is simply a mathematical statement that specifies what arithmetic operations are to be undertaken and in what order. I assure you that this is not difficult. More specifically, after stating your null and alternative hypotheses you just need to proceed through the following additional steps in the correct order:

Step 2: We must determine what the expected frequencies are. We have already accomplished this. With 10 tosses of a coin the expected frequencies, assuming the coin is not biased, are 5 heads and 5 tails.

Step 3: The equation indicates that you are to find the difference between an observed frequency and the corresponding expected frequency, $(f_o – f_e)$. In our case, there are two observed frequencies, 7 heads and 3 tails. The expected frequency for each is 5. You, therefore, calculate the first difference, which is 7 – 5. This equals 2.

Step 4: You then square the difference that you have just obtained, $(f_o – f_e)^2$. In our case, this would be 2 squared which equals 4.

Step 5: Next, you divide the squared difference that you have just calculated by its frequency expected which in this case is 5. The result of this division, $(f_o – f_e)^2 / f_e$, would be 4 / 5 which equals 0.8.

Step 6: You continue by repeating Steps 2 through 5 for each of your categories. In our example, there is only one additional category, tails. We, therefore, go back to Step 2 and confirm that the expected frequency for tails is 5. Next, in Step 3 we calculate 3 – 5 (the number of tails we observed minus the expected number of tails if the null hypothesis is correct), which equals –2. Then, as indicated in Step 4, we square –2 to obtain 4. Next, as Step 5 indicates, we divide our outcome of 4 by the expected frequency of 5 and obtain 0.8. Table 7.2 summarizes these steps.

As a check on our work, the sum of the differences between the observed and expected frequencies (obtained in Step 3) should equal 0. In our example, we have differences of +2 and –2. Their sum, as expected, is 0.

Step 7: Finally, we sum the values that we have calculated. In this case there are two categories, heads and tails, so we sum two numbers, 0.8 + 0.8 to obtain 1.6 (Table 7.2).

Congratulations! You have just calculated your first goodness-of-fit chi-square. While there are a number of steps I hope that you will agree that each is mathematically simple. Now you are ready to make your first statistical decision, to decide whether there is sufficient evidence to reject the null hypothesis that the coin is not biased (is fair).

Step 8: To complete the process, we must interpret what our chi-square value of 1.6 indicates. After all, 1.6 light years is a great many miles, but 1.6 inches is only a small distance. How do we interpret a chi-square of 1.6? In order to answer this we will need to consult the appropriate statistical table. Before doing so, however, let's reexamine the steps that we have just completed and see what they indicate.

Table 7.2 Example 1: Steps in Calculating a Goodness-of-fit Chi-square

Values	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
Heads	7	5	2	4	0.8
Tails	3	5	–2	4	0.8
			$\sum = 0$		$\sum = 1.6$

With Step 1 we state the null and alternative hypotheses.

In Step 2 we find the expected frequencies. These are derived from our null hypothesis.

In Step 3 the difference between each observed frequency and its expected frequency is calculated. Clearly, the closer our data match what is predicted from the null hypothesis, the smaller this difference will be. For instance, if we had obtained 5 heads and 5 tails in the 10 tosses, then the observed frequencies would have perfectly matched the expected frequencies and the differences would have each been zero. Alternatively, if our outcome had been 9 heads and only 1 tail, then the differences calculated in Step 3 would have been considerably greater than those that we generated.

In Step 4 we square the differences we calculated in Step 3. As a result, the outcome has a positive sign regardless of the sign of the difference that was obtained from Step 3. This was evident when we looked at our outcome of 3 tails, and from Step 3 calculated a difference of –2. This was then squared to give 4, not –4. Therefore, all of the numbers generated in Step 4 will be positive. Further, a large outcome indicates that there is a large discrepancy between our observed and our expected frequencies, while a small outcome indicates that there is only a small discrepancy.

In Step 5 we divide each category's squared difference by the expected frequency for that category. With this division we put each of the squared differences that we calculated into perspective. Fundamentally, the numerator of the chi-square equation provides a measure of how big the discrepancy is between our observed and our expected frequencies. The denominator provides a standard against which to measure this deviation. For instance, in Step 3 we found deviations of +2 and –2. Each of these outcomes was squared in Step 4, and each of these squared deviations was related to an expected frequency of 5 in Step 5. Any particular deviation is more impressive when compared against a small rather than a large standard. For instance, losing 10 pounds with a diet is more noticeable if your starting weight was 120 pounds than if it was 320 pounds.

We now understand that the outcome of Step 7, in our example this is 1.6, is necessarily positive, and we have some intuitive feel for its size. But we are still unable to conclude whether the coin the professor tossed was not biased (was fair). In order to make a decision (Step 8) we need to consult the appropriate table, in this case the chi-square table (Appendix K, Table 2). A cursory inspection of the chi-square table reveals a surprisingly large number of entries arranged into rows and columns. From previous chapters you are familiar with the distinction between Type I and Type II errors and how, as a compromise, scientists commonly set alpha at .05. In our case, we are following this convention, and thus we will be dealing with the column headed by $\alpha = .05$.

We are still left with the issue of why there are so many rows of numbers in the table, and why each row is preceded by a number associated with the two letters, df. The answer is that if there was only one chi-square distribution then there would only be a need for one row of critical values in the table. As there are many rows of values in the chi-square table this implies there is a series of chi-square distributions. Now we will explain why this is the case.

In order to calculate a chi-square value there must be at least two possible outcomes to whatever we are examining. In our example, these were either a head or tail for each toss of the coin. Clearly, if only one outcome were possible, as with a coin with two heads, then the observed frequency and the expected frequency would have to match perfectly. No matter how many times you tossed this special coin, the outcome would always be heads. But in our case of a coin with both heads and tails, there are two possible outcomes. But, can both of these outcomes vary? You might be inclined to say that as there are two outcomes that are possible, the number of heads and the number of tails, then each could vary. However, when looking at the data set as a whole we will see that there is actually only one outcome that is free to vary.

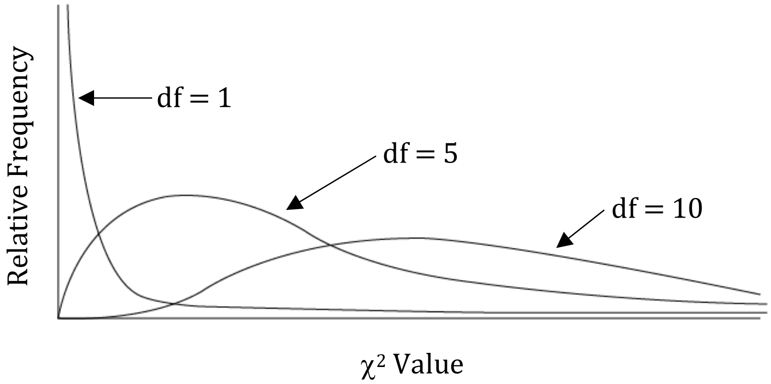
On any single toss of a coin you can obtain either a head or a tail. In our example, there were 10 tosses. And once you know that 7 of these tosses were heads, then you also know that the number of tails has to be 3. It cannot be any other number. The number of tails is not free to vary. Thus, in a study with two possible outcomes there is actually only one outcome that is free to vary. Statisticians use the term **degrees of freedom** (df) to indicate the number of outcomes that are free to vary.

Degrees of freedom (df) – The number of outcomes out of the total that are free to vary.

For the goodness-of-fit chi-square test, the degrees of freedom are equal to the number of categories possible for the outcome minus one (df = c – 1, where 'c' is the number of categories). For our example, the degrees of freedom would equal one. This is because there were two categories of the outcome (heads or tails), from which we subtract one. This is the smallest number of degrees of freedom that is possible. However, higher numbers of degrees of freedom are also possible. For instance, in tossing a die (singular of dice), there are six possible outcomes and thus there are five degrees of freedom. Once you know the number of total tosses and the number of times that five of the six sides came up, then the value of the last side is fixed. Thus, if there were a total of 10 tosses, and the numbers 1 through 5 came up eight times, then you would know that the number 6 came up two times. In other words, there are six possible outcomes, but only five degrees of freedom. Once five of the frequencies are specified, the sixth is determined.

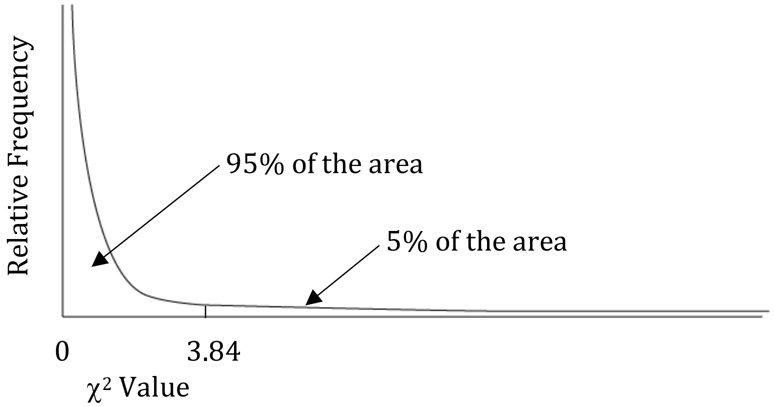
The reason degrees of freedom matter is that the shape of the chi-square distribution varies depending upon the number of degrees of freedom. This is a consequence of the mathematical equation for chi-square. In this book we will simply accept that there is not a single chi-square distribution, but rather a family of distributions with the shape of each dependent upon the number of categories of data or, more precisely, upon the number of degrees of freedom. The shapes of representative chi-square distributions are illustrated in Figure 7.2.

Figure 7.2 Shapes of Representative Chi-square Distributions



Since in our example of 7 heads and 3 tails there is only one degree of freedom, we will, for now, concentrate upon the chi-square distribution that corresponds to one degree of freedom, as is shown in Figure 7.3. The distribution illustrates the chi-square values that would be expected if the null hypothesis were in fact true. This is a theoretical distribution. It is a plot assuming that we have calculated an infinite number of chi-square values. On the X-axis are the chi-square values, on the Y-axis are the corresponding relative frequencies. It should be evident that the highest frequencies are associated with small values of chi-square, and lower frequencies are associated with larger values of chi-square.

Figure 7.3 Chi-square Distribution with 1 df



As was just stated, Figure 7.3 is an example of a theoretical distribution. No one actually collected an infinite number of chi-square outcomes. However, through a set of mathematical procedures collectively known as calculus the areas associated with different regions of this distribution have been obtained. Happily, you do not need to know any calculus to understand what is to follow. We are just going to benefit from the efforts of those who do. Thus, the situation is analogous to using a computer or a car. Most of us really do not understand how a computer or a car works, but that does not in any way preclude us from using computers and cars.

The vertical distance between a point on the X-axis and the chi-square curve indicates the relative likelihood of a particular chi-square value occurring. And the total area enclosed between the X-axis and the chi-square curve corresponds to the likelihood of all of the possible outcomes, which is 100%. Further, through calculus it has been determined that 95% of the area of the chi-square distribution with 1 df falls to the left of, and thus less than, a chi-square value of 3.84. It then follows, of course, that 5% of the area falls to the right of, or above, a chi-square value of 3.84. This latter area is given two names, the **critical region** and the **area of rejection**. In other words, if the null hypothesis is correct, then only 5% of the time will we obtain a value of chi-square by chance that is greater than 3.84 and thus so extreme that it falls in the critical region or area of rejection. The remaining 95% of the time we will obtain a value less than 3.84. Put differently, if we set alpha equal to .05, as we agreed to do previously, then we will reject the null hypothesis if the obtained chi-square value with 1 df is greater than 3.84, for 5% of the area of the distribution is above this point. Thus, 3.84 is an example of the concept of a **critical value**. By using this critical value we can be confident that when we reject the null hypothesis and accept the alternative hypothesis the likelihood of having made a Type I error is only 5%. In other words, in only 5% of the cases will we reject the null hypothesis when it is in fact correct. (This critical value of 3.84 is found in Appendix K, Table 2 by going across the row for 1 df to the first entry in the column for α = .05.)

Critical region – Area of the distribution equal to the alpha level. It is also called the Area of Rejection.

Area of rejection – Area of the distribution equal to the alpha level. It is also called the Critical Region.

Critical value – A value for a statistical test which is used to determine whether to reject or retain the null hypothesis.

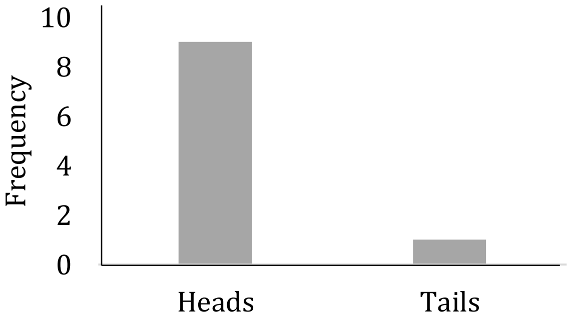
Now, going back to our example of 10 coin tosses with an outcome of 7 heads and 3 tails, you will recall that we obtained a chi-square value of 1.6. This value is less than the critical value that we have determined for 1 df, which is 3.84. Therefore, the outcome of 10 tosses with 7 heads and 3 tails does not deviate enough from the 5 heads and 5 tails that was expected, if the coin was not biased (was fair), to justify the rejection of the null hypothesis. Or, put another way, we do not conclude that the professor is using a biased coin. Instead, we accept that the outcome is simply the result of chance and thus our conclusion is to retain the null hypothesis.

It is important to recognize that we have not proven that the coin is fair (not biased). There is just insufficient evidence to conclude that it is biased. And we also recognize the we could have made a Type II error. We will discuss this more shortly.

Another Example

In order to be certain that you understand the material that has just been covered we will now do another example. In this case, let us assume that the outcome of the 10 coin tosses had been 9 heads and 1 tail. This outcome is illustrated with a bar graph in Figure 7.4.

Figure 7.4 Example 2: Bar Graph of Coin Tosses



What would you now conclude if alpha is set equal to .05? Each of the steps is shown below, and the calculations are summarized in Table 7.3.

Step 1: We start by stating the null and alternative hypotheses, and we specify our alpha level. As always, the null hypothesis is signified by H_0 and the alternative hypothesis by H_1. For this example with 10 tosses of the coin:

H_0 – The coin is not biased (it is fair); any observed deviation from the expected 50:50 outcome is due to chance.

H_1 – The coin is biased; any observed deviation from the expected 50:50 outcome is not due to chance.

We set alpha to .05.

Recall that the equation for the goodness-of-fit chi-square is

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

You proceed as in the previous example.

Step 2: You determine the expected frequencies. With 10 tosses of a fair coin the expected frequencies are 5 heads and 5 tails.

Step 3: Next you find the difference between an observed frequency and the corresponding expected frequency, $(f_o – f_e)$. The difference for the first category, heads, is 9 – 5 = 4.

Step 4: Then square the difference that you just obtained, $(f_o – f_e)^2$. This gives us $4^2 = 16$.

Step 5: The value from Step 4 is then divided by its expected frequency. This gives us $(f_o – f_e)^2 / f_e = 16 / 5 = 3.2$.

Step 6: You continue by repeating Steps 1 through 4 for the second category, tails. This obtained value is also 3.2.

As a check on your work, you then find that the sum of the differences between the observed and expected frequencies, in this case +4 and −4, is 0.

Step 7: Finally, to obtain your chi square, sum the two values you have calculated, 3.2 + 3.2 = 6.4.

Step 8: You are now ready to make your decision. Once again, we have 1 df, for there are two categories and df = c – 1. We now compare our obtained outcome of 6.4 with the critical value of chi-square with 1 df, which is 3.84 (Appendix K, Table 2). As our outcome is greater than the critical value (falls to the right of the critical value in Figure 7.3) and thus represents a highly unlikely outcome, we reject the null hypothesis and accept the alternative hypothesis. In other words, our outcome has fallen in the area of rejection, so we have sufficient grounds to conclude that it is unlikely that the coin is fair (not biased).

Table 7.3 Example 2: Steps in Calculating a Goodness-of-fit Chi-square

Values	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
Heads	9	5	4	16	3.2
Tails	1	5	−4	16	3.2
			$\sum = 0$		$\sum = 6.4$

Congratulations again! You have calculated another goodness-of-fit chi-square test and have come to another decision concerning a relationship; but be careful interpreting the outcome. You can now be *reasonably confident* that the coin is biased, but you have *not proven* that it is biased. Remember that it is possible that you have made a Type I error and have rejected the null hypothesis when in fact it is true. We indicate this by stating that we have found a statistically significant outcome. We do not say that we are certain of our decision.

The term **significant** has a very precise meaning in statistics. It simply indicates that an outcome was unlikely to have occurred by chance. In the example that we have just completed we know how unlikely, for alpha was set at .05. Thus, the probability is only 1 in 20 that, by chance alone, we could have obtained results this divergent from what was expected. We conclude, therefore, that this outcome is so unlikely to have been due solely to chance that we reject the null hypothesis that the coin is not biased (is fair) and, instead, accept the alternative hypothesis that the coin is biased. This is all that statistical significance indicates.

Significant – In statistics, the conclusion that an outcome is unlikely to have occurred by chance.

Thus, statistical significance refers to a probability. It is dealing solely with the likelihood of an outcome. By rejecting the null hypothesis we are indicating that the discrepancy between the frequencies that were observed in our sample and what would have been expected if the null hypothesis were true is unlikely to have happened by chance. Instead, we conclude that this discrepancy is likely to be indicative of a consistent quality of the coin and thus would be expected to reoccur in the future. Put another way, we have used the data from a sample (our 10 tosses of the coin) to infer a characteristic of the population (consists of all possible tosses of the coin). Thus this is an example of inferential statistics. And, our conclusion is that the coin is biased.

It is important to note that how we are now using the word significance is not what the term significance implies in our everyday conversations. When we use the term significant in a conversation we are generally interested in how meaningful or important an outcome is, not how likely or unlikely it is. No statistical procedure can completely capture how meaningful or important an outcome is, for these judgments are subjective and relative. However, later in this text we will see that there are statistical procedures that will assist you in coming to a conclusion concerning these qualities.

It is also important to remember that the statistical procedures that you are learning in this course do not ensure that you will always come to the correct conclusion. What they do instead is provide you with an accepted system for making decisions, and with this system you will know the probability of making a Type I error. For instance, in our second example we rejected the null hypothesis. Since we had specified that alpha was equal to .05, there is only a 5% chance that we made a Type I error and rejected the null hypothesis when in fact it was correct.

Progress Check

Assume you toss a coin 10 times and obtain 8 heads and 2 tails.

1. What would the value of the chi-square be?
2. How many degrees of freedom are there?
3. Would you conclude the coin is biased if alpha is set at .05?

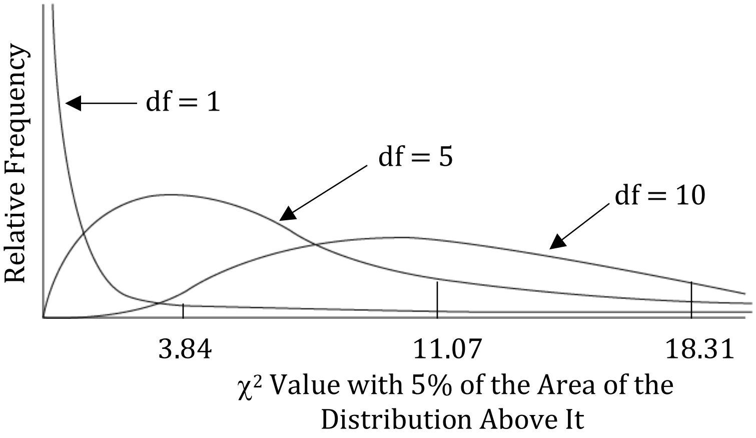
Answers: 1. 3.6 2. One 3. No

A Final Example Of The Goodness-of-fit Chi-Square

Our examples to this point have utilized one degree of freedom. However, as was noted previously, there are actually a series of chi-square distributions, each associated with its unique degree of freedom. Returning to Figure 7.2 you will note that the chi-square distributions are all positively skewed. However, the specific shape depends upon the number of degrees of freedom.

This turns out to be critically important. For instance, Figure 7.5 indicates that with one degree of freedom 5% of the area of the distribution is located to the right of a chi-square value of 3.84. Thus, in only 5% of the cases would you expect to obtain a chi-square value greater than 3.84 with one degree of freedom, if the null hypothesis is in fact true. Put another way, the odds of obtaining a chi-square value greater than 3.84 with one degree of freedom is one chance in twenty if the null hypothesis is in fact true. This also means that in 95% of the cases we would expect to obtain a chi-square value of less than 3.84 with one degree of freedom if the null hypothesis is true. This is indicated by the area to the left of the value of 3.84. The logic is exactly the same for other degrees of freedom, but with the chi-square the critical value associated with the 5% area becomes larger as the degrees of freedom increase. This is evident from an inspection of Figure 7.5. It is also evident from an inspection of the chi-square table (Appendix K, Table 2). Proceeding down the column headed by the value of α = .05, for 1 df the critical value is 3.84; for 5 df it is 11.07; and for 10 df it is 18.31.

Figure 7.5 Comparing Areas of Rejection for 1, 5 and 10 Degrees of Freedom



To illustrate, if we return to our initial example where we determined that with 10 tosses there were 7 heads but only 3 tails, you will remember that we calculated a chi-square value of 1.60. By turning to the chi-square table and using the row labeled 1 df and the column labeled .05, we find the critical value of 3.84. Our outcome of 1.60 is less than 3.84. By referring to Figure 7.3 or 7.5, we can see that our outcome is to the left of the value of 3.84 and does not fall in the region of rejection. Thus, it represents an outcome that is expected to occur more frequently than 5% of the time. In other words, we do not have sufficient evidence to warrant the rejection of the null hypothesis. Accordingly, based upon these data we tentatively accepted that the coin is fair and concluded that we have just had a string of bad luck.

You are now ready to deal with problems that have larger degrees of freedom. For instance, let us assume that you want to know if a die is fair. A die has six sides. Your null hypothesis is that for the population consisting of all tosses of the die there is no difference in the likelihood of these six outcomes. Thus, the probability is 1:6 for each outcome. The alternative hypothesis is that the probability is not 1:6 for each outcome. In other words, the alternative hypothesis is that the die is biased.

Step 1: State the null and alternative hypotheses, and specify the alpha level:

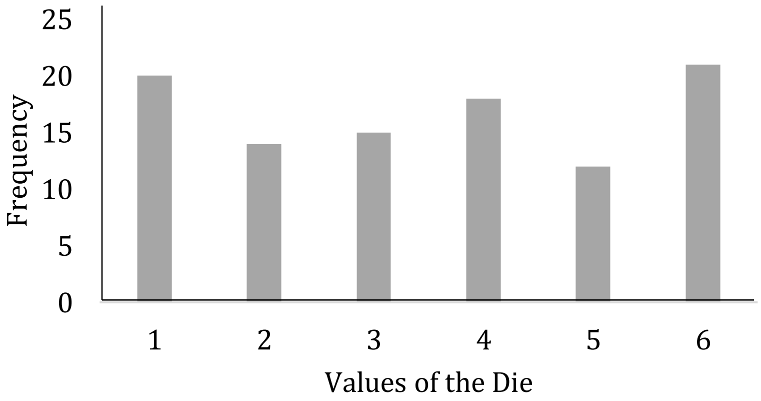
H_0 – The die is not biased (it is fair); any observed deviation from the expected 1:6 probability for each outcome is due to chance.

H_1 – The die is biased; any observed deviation from the expected 1:6 probability for each outcome is not due to chance.

Alpha is set to .05.

You then toss the die 100 times and obtain the observed frequencies in Table 7.4 (They are illustrated with a bar graph in Figure 7.6). In order to calculate the goodness-of-fit chi-square we use the same steps as previously.

Figure 7.6 Example 3: Bar Graph of Tosses of a Die



Step 2: Calculate the expected frequencies. Based upon the null hypothesis we simply divide 100, the total number of tosses, by the number of possible outcomes. This would be 100 / 6 = 16.67. Of course, 16.67 is not a frequency that could actually happen, but for the purposes of calculating the chi-square it is the expected frequency that we would use.

Step 3: Find the difference between an observed frequency and the corresponding expected frequency.

Step 4: Square the difference that was just obtained.

Step 5: Divide this squared difference by its expected frequency.

Step 6: Repeat Steps 1 through 5 for each of the additional outcomes of the die.

Step 7: Add up the six values that were just obtained. The result is 0.67 + 0.43 + 0.17 + 0.11 + 1.31 + 1.12 = 3.81. The results of these steps are illustrated in Table 7.4.

Table 7.4 Example 3: Steps in Calculating a Goodness-of-fit Chi-square

Values	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$

1	20	16.67	3.33	11.09	.67
2	14	16.67	−2.67	7.13	.43
3	15	16.67	−1.67	2.79	.17
4	18	16.67	1.33	1.77	.11
5	12	16.67	−4.67	21.81	1.31
6	21	16.67	4.33	18.75	1.12
			∑ = 0		∑ = 3.81

Step 8: Make your decision. Our degrees of freedom are equal to the number of categories minus one (df = c – 1). In this case, df would equal 6 – 1 = 5. Referring to the chi-square table (Appendix K, Table 2) and looking across the row for 5 df and down the column for $\alpha = .05$ (this is the criterion we set for rejecting the null hypothesis), we find that the critical value is 11.07. The critical value indicates that, by chance, the outcome of a chi-square test with 5 df will be greater than 11.07 only 5% of the time. Thus, in order to reject the null hypothesis with $\alpha = .05$ we would need to obtain a chi-square value greater than 11.07. Our chi-square value is 3.81, which is less than 11.07, and thus we retain the null hypothesis that the die is not biased (is fair).

If the outcome had been large enough to warrant rejecting the null, we would then examine the actual frequencies that were obtained to determine in what manner the die was biased.

Reporting The Results Of The Goodness-Of-Fit Chi-Square

If you were to report the results of the chi-square test that was just completed it would be important for the reader to be given sufficient information to fully understand the outcome. While there are a number of conventions in use, fortunately they are quite similar. In this text we will utilize the style of the American Psychological Association (APA). Until recently, you would have stated something to the effect of "the results of the testing of the die did not provide sufficient evidence to reject the null hypothesis that the die was not biased (χ^2 (5, N = 100) = 3.81, $p > .05$)". The symbol for a chi-square is χ^2. The degrees of freedom are indicated by the value 5. The n is the total number of tosses, in this case 100. The value of the obtained chi-square is 3.81 and the $p > .05$ indicates that the *probability* of this or a more extreme outcome is *greater* than the alpha level of .05 or 5%. Be careful. Our obtained chi-square value was *less* than the critical value listed in the chi-square table. Thus the probability of our, or a more extreme, outcome is *greater* than our alpha, which was set equal to .05. Our decision is, therefore, to retain the null hypothesis that the die is fair.

Purpose And Limitations Of Using The Goodness-of-Fit Chi-Square Test

1. *Uses sample data to make an inference about a population.* The goodness-of-fit chi-square is an inferential statistical procedure. The proportions obtained from a sample are being used to test an hypothesis about the proportions expected in a population. Stated differently, the goodness-of-fit chi-square is testing whether the observed frequencies differ statistically from the frequencies that would be expected if the null hypothesis is correct.
2. *Overall test of significance.* The chi-square test indicates whether a significant difference in the relative frequencies exists. In designs with more than 1 df a goodness-of-fit chi-square test with a statistically significant outcome does not indicate where the difference(s) is (are). Generally, inspection of the observed frequencies is all that is needed to indicate where the difference(s) is (are) located.
3. *No measure of effect size.* The chi-square is a test of significance. It indicates whether or not the obtained frequencies are likely to have occurred by chance if the null hypothesis is correct. With the goodness-of-fit chi-square test, no measure of effect size is commonly calculated. This concept will be explained later in the book.

Assumptions Of The Goodness-of-Fit Chi-Square Test

1. *Nominal data.* The data are in the form of frequencies or can be converted to frequencies.
2. *Observations are independent.* In other words, a subject or event is only counted once, and is not matched with or affected by another subject or event in the study.
3. *Expected frequencies cannot be too small.* For the goodness-of-fit chi-square, *the minimum acceptable size of any expected frequency is 5*. If any expected frequency is less than 5 then either an alternative statistical procedure should be utilized (refer to a more advanced statistical text) or additional data would need to be collected.

Conclusion

The goodness-of-fit chi-square is used when you have nominal data and one variable. More specifically, this procedure is used when you want to determine whether the observed frequencies obtained from a sample differ significantly from the frequencies that would be expected if the null hypothesis is true. As the chi-square test is using nominal data, or data that have been converted into nominal data, it is a nonparametric procedure. No assumptions are being made about the shape of a population distribution or the values of any population parameter, such as the mean or variability. In fact, as you know, neither a mean nor a measure of variability is calculated with nominal data.

As we have seen, the procedure to calculate the goodness-of-fit chi-square is straightforward, though there are a number of steps. It is important, however, to remember that the critical value necessary to decide whether to reject the null hypothesis varies depending upon the degrees of freedom of the study. As the course progresses you will see that this characteristic is shared with most other statistical procedures.

Glossary Of Terms

Area of rejection – Area of the distribution equal to the alpha level. It is also called the Critical Region.

Critical region – Area of the distribution equal to the alpha level. It is also called the Area of Rejection.

Critical value – A value for a statistical test which is used to determine whether to reject or retain the null hypothesis.

Degrees of freedom (df) – The number of outcomes out of the total that are free to vary.

Expected frequencies – With nominal data, the outcome that would be expected if the null hypothesis were true.

Goodness-of-fit chi-square test – An inferential procedure that tests whether observed frequencies differ from expected frequencies.

Independent – Two events, samples or variables are independent if knowing the outcome of one does not enhance our prediction of the other.

Observed frequencies – With nominal data, the actual data that were collected.

Nonparametric procedure – Statistical procedure that does not make assumptions about the population's parameters and does not assume that the population is normally distributed.

Parametric procedure – Statistical procedure that does make assumptions about the population's parameters and does assume that the population is normally distributed.

Significant – In statistics, the conclusion that an outcome is unlikely to have occurred by chance.

Questions – Chapter 7

(Answers are provided in Appendix J.)

1. Rejecting the null hypothesis when in fact it is true is a _____.
   a. Type I error
   b. Type II error
   c. Type III error
   d. Type IV error

2. The goodness-of-fit chi-square test deals with _____ data.
   a. Nominal

b. Ordinal
c. Interval
d. Ratio

3. The data in a goodness-of-fit chi-square test consist of _____.
   a. Observed frequencies
   b. Expected frequencies
   c. Degrees of freedom
   d. None of the above

4. In the statistical statement χ^2 (5, n = 100) = 3.81, p > .05, the degrees of freedom are equal to _____.
   a. 5
   b. 99
   c. 100
   d. 3.81

5. With a goodness-of-fit chi-square test, the degrees of freedom are equal to the _____.
   a. number of categories
   b. number of subjects
   c. number of categories minus one
   d. highest frequency minus two

6. If it is reported that a goodness-of-fit chi-square test has a p > .05. This indicates that the results are _____.
   a. not possible
   b. not statistically significant
   c. statistically significant
   d. not of interest

7. Practically speaking, degrees of freedom are important because they are _____.
   a. used in calculating the chi-square value
   b. used in collecting the chi-square data
   c. used in interpreting the chi-square value
   d. none of the above

8. The critical region is _____.
   a. equal to the alpha level
   b. equal to the size of beta
   c. the same as the region of rejection
   d. both 'a' and 'c'

9. If knowing that the Buffalo Bills won a football game last weekend does not aid in predicting whether they will win next weekend, statisticians would say the two events are _____.
   a. Free
   b. Independent
   c. Expected
   d. Critical

For questions 10 to 16 use the following information: suppose you toss a coin 100 times and observe 40 heads and 60 tails.

10. What is the null hypothesis?
    a. The coin is not biased.
    b. The coin is biased
    c. The observed 40:60 frequencies are <u>not</u> statistically different from the expected 50:50 ratio.
    d. The observed 40:60 frequencies are statistically different from the expected 50:50 ratio.
    e. Both a and c are correct.

11. What is the calculated value of chi-square?
    a. 1
    b. 2
    c. 3
    d. 4
    e. 5

12. How many degrees of freedom are there?
    a. 1
    b. 2
    c. 3
    d. 4
    e. 5

13. Assuming alpha is equal to .05, what is the critical value?
    a. 2.68
    b. 3.84
    c. 4.00
    d. 4.32
    e. 6.64

14. What is your decision?
    a. Accept the null hypothesis.
    b. Reject the null hypothesis.
    c. Neither accept nor reject the null hypothesis as there is insufficient data to come to a decision.

15. What is the probability of having made a Type I error?
    a. 50%
    b. 10%
    c. 5%
    d. 1%
    e. The probability of Type I error is not known.

16. What is the probability of having made a Type II error?
    a. 50%
    b. 10%
    c. 5%
    d. 1%
    e. The probability of Type II error is not known.

For questions 17 – 23 use the following information: suppose we toss a die 144 times and we observe that the number 1 occurs 36 times. (Hint – while a die has 6 sides, we have now reduced the data to just 2 categories, and we also now know that the other category, consisting of outcomes 2 through 6, must have occurred 108 times.)

17. What is the null hypothesis?
    a. The die is not biased.
    b. The die is biased
    c. The observed 36 out of 144 total tosses is *not* statistically different from the expected 1:6 ratio.
    d. The observed 36 out of 144 total tosses is statistically different from the expected 1:6 ratio.
    e. Both a and c are correct.

18. What is the calculated value of chi-square?
    a. 1
    b. 4.2
    c. 5.7
    d. 7.2
    e. 9.1

19. How many degrees of freedom are there?
    a. 1
    b. 2
    c. 3
    d. 4
    e. 5

20. Assuming alpha is equal to .01, what is the critical value?
    a. 2.68
    b. 3.84
    c. 4.00
    d. 4.32
    e. 6.64

21. What is your decision?
    a. Accept the null hypothesis.
    b. Reject the null hypothesis.
    c. Neither accept nor reject the null hypothesis as there is insufficient data to come to a decision.

22. What is the probability of having made a Type I error?
    a. 50%
    b. 10%
    c. 5%
    d. 1%
    e. The probability of Type I error is not known.

23. What is the probability of having made a Type II error?
    a. 50%
    b. 10%
    c. 5%
    d. 1%
    e. The probability of Type II error is not known.

Use the following information in questions 24 to 30. Riniolo, Koledin, Drakulic and Payne (2003) noted that 15 of 20 eyewitnesses to the sinking of the Titanic reported that the ship was breaking apart before it actually sank.

24. What is the null hypothesis?
    - a. The observed frequencies are not statistically different from the expected 50:50 ratio.
    - b. The observed frequencies are statistically different from the expected 50:50 ratio.

25. What is the calculated value of goodness-of-fit chi-square?
    - a. 1
    - b. 3.2
    - c. 5
    - d. 7.2

26. How many degrees of freedom are there?
    - a. 1
    - b. 2
    - c. 3
    - d. 4
    - e. 5

27. Assuming alpha is equal to .05, what is the critical value?
    - a. 2.68
    - b. 3.84
    - c. 4.00
    - d. 4.32
    - e. 6.64

28. What is your decision?
    - a. Accept the null hypothesis.
    - b. Reject the null hypothesis.
    - c. Neither accept nor reject the null hypothesis as there is insufficient data to come to a decision.

29. What is the probability of having made a Type I error?
    - a. 50%
    - b. 10%
    - c. 5%
    - d. 1%
    - e. The probability of Type I error is not known.

30. What is the probability of having made a Type II error?
    - a. 50%
    - b. 10%
    - c. 5%
    - d. 1%
    - e. The probability of Type II error is not known.

The text does not describe how to use of SPSS with the goodness-of-fit chi-square test as this test is not as commonly encountered as other procedures that we will be reviewing.

Chapter 8
Finding Differences with Nominal Data – II: The Chi-square Test of Independence

"Science is not a collection of facts but a way of interrogating the world."

Sharon Begley

Introduction

The chi-square statistic is not limited to analyzing frequencies obtained with a single variable, as was the case in the previous chapter. Another form of the chi-square statistic, the **chi-square test of independence**, is used when we have a design that involves nominal data and two variables. This test is underlined in Table 8.1.

Chi-square test of independence – An inferential procedure for analyzing whether the pattern of observed frequencies differs among the groups.

Table 8.1 **Overview Table of Inferential Statistical Procedures For Finding if there is a Difference**

	Type of Data			
	Nominal (Frequency)		Ordinal (Ranked)	Interval/Ratio (Continuous Measure)
Research Design		Research Design		
One Variable With At Least Two Outcomes	Goodness-of-fit Chi-Square	One IV With One Sample		One-sample z Test or One-sample t Test
		One IV With Two Or More Independent Samples	*Kruskal–Wallis H*	One-way Between–Subjects ANOVA (Only two independent samples, Independent Samples t Test)
		One IV With One Sample Having Two Or More Repeated Measures		One-way Within–Subjects ANOVA (Only two repeated measures, Dependent Samples t Test)
Two Variables, Each With At Least Two Outcomes	Chi-Square Test of Independence	Two IV Each With Two Or More Independent Samples		Two-way Between–Subjects ANOVA

Analyzing A Difference Design With Two Variables, Each With At Least Two Outcomes

In Chapter 7 we noted that in the field of statistics 'independent' has a very specific meaning for it signifies that two events, samples or variables are *not* related in a predicable fashion. The term **dependent** is used if there is a predictable relationship. Thus, if a coin is fair, then the outcome of the tosses will be independent. In other words, whether or not a head or tail was just tossed does not affect the outcome of the next flip. And if a coin is fair, then the likelihood of observing a head or tail on any one toss is 50%, regardless of what the outcomes of the previous tosses were. Unfortunately, many individuals do not understand the concept of statistical independence and instead assume that if tails has come up a number of times in a row then it is now more likely that the next toss will be a head. This is known as the **gambler's fallacy**; it has undoubtedly been responsible for the loss of a great deal of money.

Dependent – *Two events, samples or variables are dependent if knowing the outcome of one enhances our prediction of the other.*

Gambler's fallacy – *The incorrect assumption that if an event has not occurred recently, then the probability of it occurring in the future increases.*

In the scientific literature there are numerous reports of studies with nominal data using two variables. For example, Sandson, Bachna and Morin (2000) examined the relationship between Attention Deficit Hyperactivity Disorder (ADHD) and omission errors in vision. More specifically they examined on which side, the left or right, a person is less likely to see a stimulus. Neglecting to see a stimulus is known as an omission error.

The steps in conducting a chi-square test of independence closely parallel the steps that were used in Chapter 7 with the goodness-of-fit chi-square test.

Step 1: State the null and alternative hypotheses, and specify our alpha level:

H_0 – There is no difference in the distribution of omission errors between the populations of subjects diagnosed with or without ADHD.

H_1 – There is a difference in the distribution of omission errors between the populations of subjects diagnosed with or without ADHD.

As usual, we set alpha equal to .05.

In this study, each subject was assigned to either the ADHD or the no ADHD condition depending upon whether they had previously been diagnosed with ADHD. Note that the assignment for a particular subject is independent of the assignment of any other subject. The likelihood of each subject making omission errors was then assessed and, as no subject's outcome affects any others, these data are also independent.

It is important to understand that each of the 87 subjects in the study provided only a single datum (this is the singular of data, which is plural). The data, therefore, consist of joint frequencies. For instance, 36 subjects who had been diagnosed with ADHD exhibited more omissions on the right side, and 22 had more omission errors on the left side.

The data for this study are presented in Table 8.2.

Table 8.2 Example 1: Summary of the Data

		Was the Subject Diagnosed with ADHD?	
		Yes	No
Side with More Omission Errors	Right	36	25
	Left	22	4

As we have nominal data and there are two variables (ADHD diagnosis and omission side), each with two outcomes, Table 8.1 indicates that a chi-square test of independence should be employed. With two possible outcomes for each variable there are two columns and two rows of data in Table 8.2 and, consequently, this is called a 2 X 2 chi-square test of independence.

The name of this chi-square test may come as a surprise. After all, we already know that each subject's data are independent. Why, then, is the test called a test for independence? What independence is there to test for?

The chi-square test of independence examines whether the subjects diagnosed with ADHD show a different pattern of omission errors than do subjects without the diagnosis. More specifically, it is testing whether the relative frequencies of right and left omission errors differ if you were, or were not, diagnosed with ADHD. If diagnosis is not related to side of mission errors then the relative frequencies of the two groups should be similar. In that case, the outcome of the chi-square test *will not be* statistically significant and we would say the two variables are independent. However, if the relative frequencies differ enough, the outcome of the chi-square test *will be* statistically significant and we would say that the two variables are not independent. As was the case in Chapter 7, we cannot leave this decision to individual opinion. Instead, we employ an agreed-upon statistical procedure in order to come to a decision concerning whether the data indicate that a relationship exists.

The equation for the chi-square test of independence is the same as the equation we used in the previous chapter for the goodness-of-fit chi-square. And, as before, we compare the outcome we calculate with the critical value listed in the chi-square table (Appendix K, Table 2). With the chi-square test of independence there are a few more calculations involved, however none are challenging.

Step 2: As before, we begin our calculations by determining the expected frequencies. In the goodness-of-fit chi-square which was discussed in Chapter 7, the expected frequencies were a direct consequence of how the null hypothesis was stated. For instance, if the null hypothesis was that a coin was not biased (it was fair), then it followed that heads and tails would be expected to occur equally in a series of tosses. With the chi-square test of independence the situation is not quite as straightforward. Specifically, each expected frequency is calculated using the following equation:

$$\text{Expected frequency of a cell} = \frac{\text{(Frequency of its row) (Frequency of its column)}}{\text{Total n}}$$

The first issue is to understand the concept of a 'cell'. In a 2 X 2 chi-square the data consist of four frequencies. Thus, the 2 X 2 chi-square can be thought of as having four places, or cells, for the data (Table 8.2). The chi-square test will determine whether the observed pattern of frequencies in the four cells differ significantly from what would be expected by chance. As is clear from the above equation, before calculating the expected frequency for a cell it is first necessary to calculate the row totals, the column totals, and the total number of subjects. These are called marginal totals. Thus, the marginal total for the first row of our example is 36 + 25 which equals 61. Similarly, the marginal total for the first column is 36 + 22 which equals 58. For our 2 X 2 study all of the marginal totals, including the total number of subjects, are shown in Table 8.3.

Table 8.3 Example 1: Original Data with Marginal Totals

		Diagnosed with ADHD		
		Yes	No	Marginal Total
Side with More Omission Errors	Right	36	25	61
	Left	22	4	26
	Marginal Total	58	29	87

We now must calculate the expected frequency for each cell. The order in which these expected frequencies are calculated is irrelevant. However, some logical pattern should be followed so that no cell is omitted or counted twice. We will begin with the upper left cell. This cell has a row total of 61 and a column total of 58. The total number of subjects in the study is 87. Therefore, using the above equation, the expected frequency for this cell is [(61)(58)] / 87. This equals 40.67. We now calculate the expected frequency of the next cell in the first row. For this cell we would have a row total of 61 and a column total of 29. We substitute these values into the equation given above and divide by the total number of subjects in the study to obtain the expected frequency of 20.33. We then proceed with the two cells of the second row. The results are shown in Table 8.4.

Table 8.4 Example 1: Expected Frequencies

		Diagnosed with ADHD	
		Yes	No
	Right	40.67	20.33
Side with More Omission Errors			
	Left	17.33	8.67

Steps 3 - 7: We could then proceed by constructing a table that incorporates the steps needed to calculate a chi-square test of independence (Table 8.5). This table is identical to the table used in calculating the value of the goodness-of-fit chi-square and utilizes steps 3 - 7 described in Chapter 7.

Table 8.5 Example 1: Steps in Calculating a Chi-square Test of Independence

Cells	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
1	36	40.67	–4.67	21.81	0.54
2	25	20.33	4.67	21.81	1.07
3	22	17.33	4.67	21.81	1.26
4	4	8.67	–4.67	21.81	2.52
			$\sum = 0$		$\sum = 5.39$

Alternatively, you may find it is more efficient to directly calculate the chi-square:

$$\chi^2 = \Sigma \frac{(\text{Frequency observed – Frequency expected})^2}{\text{Frequency expected}}$$

This can be written as:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

For our example:

$$\chi^2 = \frac{(36 - 40.67)^2}{40.67} + \frac{(25 - 20.33)^2}{20.33} + \frac{(22 - 17.33)^2}{17.33} + \frac{(4 - 8.67)^2}{8.67}$$

$$= 0.54 + 1.07 + 1.26 + 2.52$$

$$= 5.39$$

The outcome is the same with either approach.

Step 8: As with the goodness-of-fit chi-square, we must now consult the chi-square table (Appendix K, Table 2) in order to compare our outcome with the critical value. In order to do so we must first determine our degrees of freedom. For the chi-square test of independence,

Degrees of freedom (df) = (Number of rows – 1)(Number of columns – 1)

For our example, since we have 2 rows and 2 columns, we have df = (2 – 1)(2 – 1) which equals 1 X 1 or 1. (Having 1 df for a 2 X 2 chi-square is logical. If the marginal totals are known, then only one cell frequency is free to vary. Once any cell frequency is chosen, the other three cell frequencies are fixed.) With alpha equal to .05, the critical value found in the chi-square table (Table 2 in Appendix K) is 3.84. As our obtained chi-square, 5.39, is larger than the critical value, we reject the null hypothesis that the two samples came from populations with the same proportions for side of omission, and accept the alternative hypothesis that the samples came from populations with proportions that differed for side of omission. Specifically, inspection of Table 8.2 indicates that the pattern of frequencies for individuals with and without the ADHD diagnosis differed – the individuals without the ADHD diagnosis were less likely to omit stimuli presented to the left side than were those individuals with the ADHD diagnosis.

It may be helpful to review what we have just accomplished. We began with two samples and utilized a chi-square statistic to come to a decision concerning the populations from which they were drawn. The procedure we used has several advantages over simply looking at the data and jumping to a conclusion. First, the steps are agreed upon and thus others will proceed as we did and will come to the same decision. Personal opinion is *not* the basis for our conclusion. Second, by setting alpha at a particular value, in this case .05, we have defined the magnitude of our Type I error rate. Remember, there is no guarantee that we made the correct decision when we rejected the null hypothesis. It is possible that, by chance, we obtained two very unlikely samples. At least we know, though, that the probability of having rejected the null hypothesis when in fact it was true is only .05 or 5% since this was the value we chose for our alpha level.

We would report our finding in a journal article in the same way that we would report a goodness-of-fit chi-square, giving the df, the number of subjects, the calculated chi-square value, and whether the probability of the outcome is less than or greater than the alpha level. Based upon our calculations we would report χ^2 (1, N = 87) = 5.39, $p < .05$. However, as you will see, with SPSS we can obtain a more accurate determination for the chi-square. And SPSS provides a **p-value**, the probability of our, or a more extreme, outcome occurring by chance assuming the null hypothesis is correct. Thus we would report (χ^2 (1, N = 87) = 5.38, $p = .020$). Note that the p-value of .020 is less than our α of .05, confirming that we would reject the null hypothesis.

<u>p-value</u> – The probability of an outcome, or a more extreme outcome, occurring by chance assuming the null hypothesis is correct. To be statistically significant, the p-value

must be less than the alpha level, which is usually .05.

An Important Observation

The chi-square test of independence has thus far been discussed as a procedure for determining whether the distribution or pattern of frequencies observed for each group differ. Stated another way, we have been testing whether a difference observed in the pattern of the observed frequencies is likely to generalize to the corresponding populations. It is important to note, however, that with nominal data the distinction between studies looking for a difference and studies looking for an association is often not as clear as with designs utilizing ordinal, interval or ratio data. Consequently, the chi-square test of independence is also commonly used with studies examining whether there is likely to be an association between the variables in the corresponding populations (underlined in Table 8.6). Thus, the chi-square test of independence is also often called the **chi-square test of association**. For example, it was reported in the *Buffalo News* that if the quarterback of the Buffalo Bills threw two or more interceptions in a game the win-loss ratio was a disappointing 1 to 13. In contrast, in those games in which there were none or only one interception, the win-loss ratio improved to 8 to 8. These data would often be analyzed with a chi-square test of independence. And it would be appropriate to say either that there was a difference in the pattern of wins and losses, or that there was an association between the number of interceptions and the likelihood of losing the game. (These are two ways of saying the same thing.) However, due to the lack of random assignment of subjects or experimental control of an IV, it would not be appropriate to say that throwing interceptions caused the team to lose.

Chi-square test of association – Another name for the chi-square test of independence.

Table 8.6 Overview Table of Procedures For Finding if there is an Association

	Type of Data		
	Nominal (Frequency)	Ordinal (Ranked)	Interval/Ratio (Continuous Measure)
Research Question			
Association:	Chi-Square Test of Independence		
Correlation:	*Phi r*[a]	*Spearman r*[b]	Pearson r *Multiple Correlation*[c]
Regression:			Regression *Multiple Regression*[c]

Italicized items are reviewed in the following appendixes:

a. Appendix B
b. Appendix C
c. Appendix D

A Second Example

Another example may be helpful to be certain that you understand how to calculate a chi-square test of independence, and to show that with nominal data the same study can be analyzed as either focusing upon a difference or an association. It has been reported that men are generally more distressed by the sexual infidelity than the emotional infidelity of their partners, whereas women are more distressed by the emotional infidelity than the sexual infidelity of their partners. Mathes (2003) re-examined this issue. His results are summarized in Table 8.7.

Table 8.7 Example 2: Summary of the Data for the Second Example

	Women	Men
More distressed by emotional infidelity	42	12
More distressed by sexual infidelity	17	48

Each of the steps will be shown below, but it is strongly suggested that you try this example on your own and use the text as a check on the accuracy of your work. At the end of this example we will discuss what statistical significance indicates as well as what it does not indicate.

Step 1: We begin by stating the null and alternative hypotheses, and we specify our alpha level:

H_0 – There is no difference in the distribution of answers for women and men. (There is no association between the answers of men and women.)

H_1 – There is a difference in the distribution of answers for women and men. (There is an association between the answers of men and women.)

As usual, we have set our alpha level to .05.

Step 2: The next step is to determine the expected frequencies using the following equation:

$$\text{Expected frequency of a cell} = \frac{\text{(Frequency of its row) (Frequency of its column)}}{\text{Total n}}$$

Before we can calculate the expected frequency for each cell it is first necessary to calculate the row totals, the column totals, and the total number of subjects. For our 2 X 2 study, these marginal totals are indicated in Table 8.8.

Table 8.8 Example 2: Original Data with Marginal Totals

	Women	Men	Marginal Total
More distressed by emotional infidelity	42	12	54
More distressed by sexual infidelity	17	48	65
Marginal Total	59	60	119

We now can calculate the expected frequency for each cell, which is given by [(Row total)(Column total)] / Total n. Thus, for the upper left cell the expected frequency would be [(54)(59)] / 119. This equals 26.77. The same procedure would be followed to find the remaining three expected frequencies. The result of these calculations is shown in Table 8.9.

Table 8.9 Example 2: Expected Frequencies

	Women	Men
More distressed by emotional infidelity	26.77	27.23
More distressed by sexual infidelity	32.23	32.77

Steps 3 - 7: We could then proceed by constructing a table summarizing the calculations needed to obtain a value for the chi-square test of independence (Table 8.10).

Table 8.10 Example 2: Steps in Calculating a Chi-square Test of Independence

Cells	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
1	42	26.77	15.23	231.95	8.66
2	12	27.23	–15.23	231.95	8.52
3	17	32.23	–15.23	231.95	7.20
4	48	32.77	$\underline{15.23}$	231.95	$\underline{7.08}$
			$\sum = 0$		$\sum = 31.46$

Alternatively, we could determine the value of the chi-square by calculating [(Frequency observed – Frequency expected)^2 / Frequency expected] directly for each of the four cells, and then adding the results together:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(42 - 26.77)^2}{26.77} + \frac{(12 - 27.23)^2}{27.23} + \frac{(17 - 32.23)^2}{32.23} + \frac{(48 - 32.77)^2}{32.77}$$

$$= 8.66 + 8.52 + 7.20 + 7.08$$

$$= 31.46$$

The outcome is the same using either method.

Step 8: We make our decision. To do so we must find our degrees of freedom. For a 2 X 2 chi-square the degrees of freedom would equal (2 – 1)(2 – 1), or 1 X 1 = 1. Referring to the chi-square table (Table 2 in Appendix K) we find that the critical value for 1 df with alpha set at .05 is 3.84. As our obtained chi-square of 31.46 is greater than this critical value we reject the null and accept the alternative hypothesis. In fact, a chi-square of 31.46 has a likelihood of happening, by chance, of less than .001, or 1 chance in 1000. We recognize, nevertheless, that we may have made a Type I error, and thus have not proven that there is a difference in the patterns of the proportions between the two populations.

Researchers would indicate that the calculated chi-square was substantially greater than the critical value of 3.84 by showing that the probability of the outcome was considerably less than .05. In a journal, the outcome would be reported as $\chi^2 = (1, N = 119) = 31.46, p < .001$. As you will see shortly, use of a computer package such as SPSS permits a more accurate report.

We have just completed our second chi-square test of independence. Hopefully you agree that there is nothing particularly challenging about analyzing data with this statistical test. It is, of course, important to proceed through each step in a careful manner, but no step is mathematically or conceptually difficult.

Progress Check

1. With the goodness-of-fit chi-square the expected frequencies are determined from the _____, whereas with a chi-square test of independence they must be _____.
2. With the chi-square test of independence, if the pattern of the frequencies is similar the outcome will _____ statistically significant.
3. The equation for the chi-square test of independence is _____ as the equation for the goodness-of-fit chi-square.

Answers: 1. null hypothesis; calculated 2. not be 3. the same

This is a good time to examine in more detail what we have and have not found. In our just-completed chi-square test of independence, our calculated outcome was greater than the critical value that we obtained from the chi-square table (Table 2 in Appendix K). Thus we concluded that there was a statistically significant difference between the genders (there was an association) for what causes distress within a relationship. This indicates that the pattern of the observed frequencies was unlikely to have occurred by chance. However, it is important to understand how statistical significance in a chi-square is dependent upon two characteristics of the data, (1) the distribution of the observed frequencies and (2) the sample size.

In order to see how sample size affects the chi-square, we will, for illustration purposes, simply double each of the numbers in Table 8.7, which also doubles the marginal totals. This is illustrated in Table 8.11. By doing so, we have not changed the pattern of the observed frequencies; we have simply doubled the size of the study. Please check that the new value of the chi-square would be 62.87. Thus, while the pattern of the frequencies within the chi-square tables (Tables 8.7 and 8.11) has stayed the same, as has the degrees of freedom, the magnitude of the calculated chi-square has doubled! (If you do the calculations, you will notice that there is a very minor discrepancy due to rounding error.) This means that we now have an even less likely outcome than was previously found. Why would this be the case? Actually, a moment's thought will confirm that this is exactly what one would expect to happen. The chi-square test of independence indicates the likelihood that an outcome would occur *by chance* if the null hypothesis were true. Though the obtained relative frequencies have remained the same, the new outcome is less likely because it is now based upon twice the data. You would be more confident that you are a good student in a course if you have obtained four 'A' grades instead of just two. The chi-square is influenced in the same manner – the magnitude of the chi-square increases if it is based upon more data.

Table 8.11 Illustration of Data Being Doubled

	Women	Men	Marginal Total
More distressed by emotional infidelity	84	24	108
More distressed by sexual infidelity	34	96	130
Marginal Total	118	120	238

If you have followed the logic of the argument then it should be clear that this raises a problem for how we interpret a significant statistical outcome. We have found that the size of the calculated chi-square is affected by the size of the set of data in the study. It follows, then, that while a significant outcome could occur from having a large effect with a moderately-sized data set, as in our example (Table 8.7), it could also occur by either having a very large effect in a small data set, or by having only a small effect in a very large data set. Therefore, a statistically significant chi-square based upon a small data set may actually be more impressive than a significant chi-square that is based on a much larger sample! This is true because to have a statistically significant outcome with only modest sample sizes indicates that your independent variable must have had a large effect, and having a large effect is closer to what we generally mean by the word 'significance' in our everyday conversations. Fortunately, there is a statistical measure of **effect size** for use with a 2 X 2 chi-square test; it is the phi-coefficient (pronounced fie), which is usually just called **phi**. Phi is the Greek letter ϕ. [The relationship between the chi-square test of independence, the phi-coefficient, and the phi correlation (phi r) is discussed in Appendix B.] A measure of effect size is usually only calculated after a statistically significant outcome has been found. Assuming that your study had sufficient power (reviewed in Appendix E) in most cases it would not make sense to find a measure for how strong an effect is unless you first found evidence that the effect existed. Fortunately, the phi-coefficient can easily be obtained once a statistically significant 2 X 2 chi-square has been calculated, as is evident with the following equation:

$$\text{Phi} = \phi = \sqrt{\frac{\chi^2}{\boldsymbol{n}}}$$

<u>Effect size</u> – A measure of how 'strong' a statistically significant outcome is.

<u>Phi</u> – Measure of effect size for the 2 X 2 chi square tests of independence.

A phi-coefficient with a value of approximately .50 or larger is considered to be indicative of a large effect, a phi-coefficient of approximately .30 is considered to be indicative of a medium-sized effect and a value of approximately .10 is indicative of a small effect (Table 8.12).

Table 8.12 Interpretation of Phi

Small Effect	Medium Effect	Large Effect
.10	.30	.50

One of the advantages of the phi-coefficient is that, unlike the chi-square, it is not affected by the sample size. You can confirm this by calculating the phi-coefficient based upon the data presented in Table 8.7 and calculating it again for the doubling of these data, which is illustrated in Table 8.11. As the following calculations show, the two phi-coefficients are identical:

$$\phi = \sqrt{\frac{31.46}{119}}$$
$$= .51$$

$$\phi = \sqrt{\frac{62.87}{238}}$$
$$= .51$$

The pattern of the observed frequencies, which are indicative of the effect sizes, remained the same within the two tables and thus the phi-coefficients remained the same. Our calculated phi-coefficient of .51 indicates a large-sized effect.

Reporting The Results Of A 2 X 2 Chi-Square Test Of Independence

Journals commonly require that a measure of effect size be reported along with the results of a test of significance. For the data in Table 8.7 we would state that the null hypothesis was rejected and that the effect size was large ($\chi^2 = (1, N = 119) = 31.46, p < .001, \phi = .51$). We would then state that women are more distressed by the emotional infidelity of their partners whereas men are more distressed by their partners' sexual infidelity. With a statistical package such as SPSS we are able to more accurately indicate our findings and would write ($\chi^2 = (1, N = 119) = 31.45, p < .001, \phi = .51$). Note how close our calculations are to what is found with SPSS.

Review

Before proceeding with a more complex design using the chi-square test of independence it might be good to pause and review what we have covered in Chapters 7 and 8. In Chapter 7 we introduced the chi-square goodness-of-fit test. This test is utilized when there are frequency (nominal) date. More specifically, it examines whether there is a statistically significant difference between the observed and expected frequencies when there is one variable. Chapter 8 introduced the chi-square test of independence. It is also used with frequency (nominal) data, but now there are two variables. We began our introduction to the chi-square test of independence as a continuation of our discussion of the procedures used with difference designs (Table 8.1). We then noted that the distinction between difference and association designs is not as evident with nominal data as it is with ordinal, interval or ratio data, and that the chi-square test of independence can also be utilized with association designs (Table 8.6). As a consequence we learned that this test is also called the chi-square test of association.

There has also been a discussion of degrees of freedom, how to determine a critical value from a statistical table and why it is important to also report the effect size, as measured by the phi-coefficient. You will see that the same general approach applies to the statistical tests utilized with ordinal as well as interval or ratio data.

A Third Example: This Time of a Larger Chi-Square

We will now continue our exploration of the chi-square test of independence by turning to a more complex example. Fortunately, much of what you have just learned about analyzing data with a 2 X 2 chi-square carries over to situations which require larger chi-squares. However, as you will see there are also important differences (Table 8.13).

Table 8.13 Comparison of 2 X 2 Chi-Square with Larger Chi-Squares

Chi-square	Equation	Effect Size	Post Hoc Test
2 X 2	$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$	Phi	None
Larger than 2 X 2	Same	Cramer's V	Bonferroni Method

We conclude our discussion of the chi-square test of independence with a review of a study by Perfect (2003). He examined factors that might influence the accuracy of witness identification of suspects in a police lineup. Previous work had found, surprisingly, that if an eyewitness described a perpetrator they were subsequently *less* accurate in picking the perpetrator out of a subsequent lineup. The study by Perfect (2003) tested undergraduates to determine how different types of intervening activity would affect their subsequent success with a lineup. For the control condition, the intervening activity was simply reading a magazine article for 10 minutes. One experimental condition was engaging in a task that required concentration on the details of a series of stimuli. A second experimental condition was engaging in a task that required concentration on more global aspects of stimuli. Note that in this study each condition was independent; no subject was in more than one condition, nor were the subjects in one condition related in any way with the subjects in the other two conditions. Further, the data are nominal; each subject was either successful or not successful with the lineup identification.

Step 1: State the null and alternative hypotheses, as well as the alpha level:

H_0 – There is no difference between the conditions in the distribution of successful and not successful answers. (There is no association between type of task and eyewitness success.)

H_1 – There is a difference between the conditions in the distribution of successful and not successful answers. (There is an association between type of task and eyewitness success.)

In this study, alpha was set at .05.

An advantage of all chi-square procedures is that the data can be represented simply. The data for the Perfect (2003) study are shown in Table 8.14.

Table 8.14 Example 3: Summary of the Data

	Experimental Condition		
	Control	Detail	Global
Successful ID	21	24	13
Not successful ID	9	6	17

As there are two rows (each undergraduate was either successful or not successful with their identification) and 3 columns (each undergraduate was assigned to one of three conditions), this is a 2 X 3 design. More specifically, as there are nominal data we would utilize a 2 X 3 chi-square test of independence.

Step 2: The expected frequencies are found by using the same equation as for a 2 X 2 chi-square:

$$\text{Expected frequency of a cell} = \frac{\text{(Frequency of its row) (Frequency of its column)}}{\text{Total n}}$$

Once again, before calculating the expected frequency for a cell it is first necessary to calculate the row totals, the column totals and the total number of subjects. For our 2 X 3 study, these marginal totals are indicated in Table 8.15.

Table 8.15 Example 3: Original Data with Marginal Totals

	Control	Detail	Global	Marginal Total
Successful ID	21	24	13	58
Not successful ID	9	6	17	32
Marginal Total	30	30	30	90

We now must calculate the expected frequency for each cell. Beginning with the upper left cell we have a row total of 58 and a column total of 30. The total number of subjects in the study is 90. Therefore, using the above equation the expected frequency for this cell is [(58)(30)] / 90, which equals 19.33. We now calculate the expected frequency of each of the other cells. The results are shown in Table 8.16.

Table 8.16 Example 3: Expected Frequencies

	Control	Detail	Global
Successful ID	19.33	19.33	19.33
Not successful ID	10.67	10.67	10.67

Steps 3 - 7: We could then proceed to calculate the value of the chi-square by constructing a table as we did previously (Table 8.17).

Table 8.17 Example 3: Steps in Calculating a Chi-square

Cells	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
1	21	19.33	1.67	2.79	0.14
2	24	19.33	4.67	21.81	1.13
3	13	19.33	–6.33	40.07	2.07
4	9	10.67	–1.67	2.79	0.26
5	6	10.67	–4.67	21.81	2.04
6	17	10.67	6.33	40.07	3.76
			$\sum = 0$		$\sum = 9.40$

Alternatively we could directly calculate our chi-square:

$$\chi^2 = \sum \frac{(\mathrm{f_o} - \mathrm{f_e})^2}{\mathrm{f_e}}$$

$$= \frac{(21 - 19.33)^2}{19.33} + \frac{(24 - 19.33)^2}{19.33} + \frac{(13 - 19.33)^2}{19.33} + \frac{(9 - 10.67)^2}{10.67} + \frac{(6 - 10.67)^2}{10.67} + \frac{(17 - 10.67)^2}{10.67}$$

$$= 0.14 + 1.13 + 2.07 + 0.26 + 2.04 + 3.76$$

$$= 9.40$$

The outcomes are identical.

Step 8: We now consult the chi-square table (Appendix K, Table 2) in order to compare our outcome with the critical value. In order to do so we must determine our degrees of freedom. For the chi-square test of independence:

df = (Number of rows – 1)(Number of columns – 1)

For our example, since we have 2 rows and 3 columns, we have df = (2 – 1)(3 – 1) which equals 1 X 2, or 2. With alpha equal to .05, the critical value, found in the chi-square table (Table2 in Appendix K), is 5.99. As our obtained chi-square, 9.40, is larger than the critical value we reject the null hypothesis that the type of task does not affect eyewitness success, and accept the alternative hypothesis that the type of task does affect eyewitness success. We could instead state that the samples came from populations with different proportions, and thus the two variables are not independent. Or, we could note that there is an association between type of task and eyewitness success.

We are still faced with two issues. First, we have not yet calculated a measure for effect size. Second, the chi-square procedure provides an overall test of significance for the entire study but does not indicate where the significant difference(s) is (are). With a 2 X 2 chi-square this is not an issue for there are only two conditions and thus, if there is a significant outcome, the difference has to be between the two categories. With three or more conditions the issue is not so clear. For instance, in our case there are three conditions (Control, Detailed and Global). The significant difference(s) in the obtained proportions could be between a pair of conditions (between condition 1 and condition 2, between condition 1 and condition 3, or between condition 2 and condition 3), any two of these comparisons, or all three of these comparisons. In addition, more complex comparisons could be involved, such as condition 1 versus a combination of conditions 2 and 3, or condition 2 versus a combination of conditions 1 and 3, or condition 3 versus a combination of conditions 1 and 2. In this book we will deal only with what are called pairwise comparisons, the comparisons involving the initial conditions, not any of the more complex combinations. The chi-square test that we have just completed does not specify which of these outcomes is statistically significant. It simply indicates that at least one comparison within the data is expected to be significant. We will examine the issue of effect size first, and then describe a procedure for specifying where a difference(s) within a significant chi-square is (are) located.

With a 2 X 2 chi-square, we have seen that the phi-coefficient provides a measure of effect size. Fortunately, only a minor modification is required to calculate a measure of effect size for any size chi-square. This measure of effect size is called **Cramer's V**:

Cramer's V = $\surd \frac{\chi^2}{n(df)}$

where df = the ***smaller*** of (r – 1) or (c – 1)

<u>Cramer's V</u> – Measure of effect size for chi square tests of independence larger than 2 X 2.

This equation is very similar to the equation provided earlier for the phi-coefficient. In fact, the only difference is the inclusion of the degrees of freedom. (However, note that the definition of degrees of freedom has changed, it is *not* the same as the definition used with the overall chi-square. Also, note that in the situation where there is one df, the phi-coefficient and Cramer's V are identical. In our case the df for Cramer's V = 1 as the smaller of the rows – 1 or columns – 1 is equal to 2 – 1 = 1.)

We calculate Cramer's V as follows:

Cramer's V = $\surd \frac{9.4}{(90)(1)}$

= $\surd \frac{9.4}{90}$

= 0.32

The interpretation of Cramer's V is slightly more complex than the interpretation was for the phi coefficient. As Table 8.17 indicates, the interpretation of the effect size will vary depending upon the degrees of freedom that were used in the calculation of Cramer's V (Cohen, 1988). Note that the degrees of freedom used in Table 8.18 are for Cramer's V and are *not* necessarily the same as the degrees of freedom from the overall chi-square.

Table 8.18 Interpretation of Cramer's V

Cramer's V df	Small Effect	Medium Effect	Large Effect
1	.10	.30	.50
2	.07	.21	.35
3	.06	.17	.29
4	.05	.15	.25
5	.04	.13	.22

By checking Table 8.18, you will see that, with 1 df the interpretation of Cramer's V is the same as was previously given for the phi-coefficient. This is what one would expect, for they provide the same outcome when there is 1 df. In our case, we obtained a Cramer's V of .32 with 1 df, which would be a medium-sized effect.

We now turn to our second question: How do we identify the specific samples that differ when a chi-square with more than two conditions has been found to be statistically significant? What we are dealing with here are called **post hoc comparisons**. Post hoc comparisons are employed when the overall test of significance involves more than two conditions or samples. (In Chapter 12 we will expand this definition.) From Table 8.1 (or Appendix L) you will see that studies with more than two conditions (samples) can occur not only with chi-square designs but also following a significant Kruskal-Wallis H test when there are ordinal data (reviewed in Appendix A), and with ANOVAs when there are interval or ratio data.

<u>Post hoc comparisons</u> – Statistical procedures utilized following an initial, overall test of significance in order to identify the specific conditions (samples) that differ.

In our current example we have a statistically significant 2 X 3 chi-square. If it were not significant we would not conduct any post hoc test. However, since it is significant and we want to know where the difference(s) is (are), we would conduct every pairwise comparison. With three conditions (2 df in the overall chi-square), there are three possible pairwise comparisons. As we noted previously these are between condition 1 and condition 2; between condition 1 and condition 3; and between condition 2 and condition 3. Inspection of Table 8.1 indicates that with these data we would compute three additional 2 X 2 chi-square statistics, each testing one of the above comparisons. Thus, one 2 X 2 chi-square would compare the proportions obtained from the 'Control' condition with the proportions obtained from the 'Detail' condition. Another would compare the proportions obtained from the 'Control' condition with the proportions obtained from the 'Global' condition. The final chi-square would compare the proportions obtained from the 'Detail' condition with the proportions obtained from the 'Global' condition.

There is a potential problem, however, when you conduct multiple comparisons. If you keep alpha equal to .05 for each comparison then when you conduct a large number of comparisons you increase the likelihood of finding statistical significance when, in fact, there is no relationship in the populations. Here is why. With alpha set at .05 for a comparison you know there is a 5% chance of making a Type I error. In other words, with this one comparison there is one chance in twenty that you will reject the null hypothesis when in fact it is correct. But what happens if you conduct a series of statistical comparisons, each with their alpha set at .05? Clearly, since each comparison has one chance in twenty of leading to a Type I error, if you conduct numerous comparisons it will become increasingly likely that you will commit at least one Type I error. The problem is that alpha is being set per comparison. Earlier we found that studies with very large sample sizes may lead to statistically significant, but potentially meaningless outcomes, and you learned that by using the phi-coefficient or Cramer's V we can interpret the effect size independent of the sample size. As you might expect there is also a solution for the problem of increased error rate that arises from conducting more than one post hoc comparison.

One of the easiest methods is to divide the overall alpha rate that you want to maintain by the number of post hoc comparisons you will make. In our example, as usual, the overall alpha was set at .05. We would then divide .05 by the number of post hoc chi-square tests that we will conduct, and use this more stringent requirement when determining our critical values. By doing so we maintain the overall, or experimentwise, error rate at .05. This is known as the **Bonferroni method** of controlling the error rate.

> *Bonferroni method – A procedure to control the Type I error rate when making numerous comparisons. In this procedure the alpha level that the experimenter has set is divided by the number of comparisons.*

In the current example we would conduct all three of the possible pairwise comparisons, as shown below. And using the Bonferroni method we would divide our initial alpha of .05 by three, since we are making three comparisons. Thus, we would be using the critical value associated with an alpha equal to .05 / 3 = .0167. Since the chi-square table (Appendix K, Table 2) that we have used previously does not include this particular alpha we would need to turn to a more extensive table, or utilize a computer program. For an alpha of .0167 with 1 df the critical value is 5.73. If the Bonferroni method had not been utilized our critical value, with 1 df, would have been 3.84. Let us see what difference this makes.

For each of the three post hoc comparisons the expected frequencies will need to be re-calculated.

The data and marginal totals for the comparison between the 'Control' and 'Detail' conditions are shown in Table 8.19.

Table 8.19 Post Hoc Comparison for the Control and Detail Conditions: Original Data with Marginal Totals

	Control	Detail	Marginal Total
Successful ID	21	24	45
Not successful ID	9	6	15
Marginal Total	30	30	60

The expected frequencies are shown in Table 8.20.

Table 8.20 Expected Frequencies for the Post Hoc Comparison of the Control and Detail Conditions

	Control	Detail
Successful ID	22.5	22.5
Not successful ID	7.5	7.5

You should confirm that the chi-square value for this comparison is 0.80.

The data and marginal totals for the comparison between the 'Control' and 'Global' conditions are shown in Table 8.21.

Table 8.21 Post Hoc Comparison for the Control and Global Conditions: Original Data with Marginal Totals

	Control	Global	Marginal Total
Successful ID	21	13	34
Not successful ID	9	17	26
Marginal Total	30	30	60

The expected frequencies are shown in Table 8.22.

Table 8.22 Expected Frequencies for the Post Hoc Comparison of the Control and Global Conditions

	Control	Global
Successful ID	17	17
Not successful ID	13	13

You should confirm that the chi-square value for this comparison is 4.34.

Finally, the data and marginal totals for the comparison between the 'Detail' and 'Global' conditions are shown in Table 8.23.

Table 8.23 Post Hoc Comparison for the Detail and Global Conditions: Original Data with Marginal Totals

	Detail	Global	Marginal Total
Successful ID	24	13	37
Not successful ID	6	17	23
Marginal Total	30	30	60

The expected frequencies are shown in Table 8.24.

Table 8.24 Expected Frequencies for the Post Hoc Comparison of the Detail and Global Conditions

	Detail	Global
Successful ID	18.5	18.5
Not successful ID	11.5	11.5

You should confirm that the chi-square value for this comparison is 8.54.

Note that each of these three 2 X 2 chi-square tests has one degree of freedom. Therefore, two of these three post hoc comparisons would be statistically significant at the .05 level, *if the Bonferroni method were not used*, as two of the outcomes are greater than the critical value of 3.84. However, note that only one of our three post hoc tests met the more conservative criterion of 5.73 set by the Bonferroni method. When reporting our finding, we use the 5.73 criterion, but report p < .05 since with the Bonferroni method this is what we set as the 'Experimentwise' Type I error rate.

Reporting The Results Of A Chi-Square Test Of Independence Larger Than 2 X 2

For the data in the overall 2 X 3 chi-square in Table 8.14 we would report that the null hypothesis was rejected. This indicates that the intervening activity affected the rate of successful identification. And we would include our measure of effect size. Based upon our calculations we would indicate this by writing (χ^2 (2, N = 90) = 9.40, $p < .01$, *Cramer's V* = .32). We can indicate the chi-square value more precisely and include a p-value by using a statistical packages such as SPSS (χ^2 (2, N = 90) = 9.41, $p = .009$, *Cramer's V* = .32). (Note that the p-value of .009 is less than our α of .05, confirming that we would reject the null hypothesis.) This would be followed by a statement indicating that three post hoc pairwise comparisons, using the Bonferroni method, were then conducted and only the comparison of the 'Detail' condition with the 'Global' condition was found to be significant (χ^2 (1, N = 60) = 8.54, $p < .05$). With these statements we have provided the reader with a great deal of information. We indicated that we conducted an overall chi-square test for an independent samples experiment; we told the reader the number of degrees of freedom in the design, as well as the number of participants in the study; and we indicated the value of the chi-square and noted that it was statistically significant. Further, we provided a measure of effect size so that the readers can judge the strength of the relationship. The readers are thus in a position to make an informed decision about how meaningful the outcome is. Finally, we indicated where, within the study, this significant effect occurred. All of this was communicated efficiently, using a minimum number of words.

Purpose And Limitations Of The Chi-Square Test Of Independence (Chi-Square Test of Association)

1. *It is a test for whether there is a difference (or an association) among the proportions.* The null hypothesis is that the observed frequencies are distributed similarly within each of the populations. In other words, the relative frequencies are expected to be the same for each of the conditions and thus any difference in the pattern of observed proportions is due to chance.
2. *It provides an overall test of significance.* In designs that are larger than 2 X 2, a statistically significant outcome indicates that a difference in the relative frequencies exists between the conditions, but the overall chi-square test does not indicate where the difference(s) is (are). Subsequent post hoc chi-square tests are conducted to identify the specific conditions that differ.
3. *The test does not provide a measure of effect size.* The chi-square is a test of statistical significance. It indicates whether or not an outcome is likely to have occurred by chance if the null hypothesis is correct. If the chi-square statistic is significant, a measure of effect size, phi or Cramer's V, should then calculated.

Assumptions Of The Chi-Square Test Of Independence

The assumptions of the chi-square test of independence (chi-square test of association) are similar to those of the goodness-of-fit chi square.

1. *Nominal data.* The data are in the form of frequencies or can be converted to frequencies.
2. *Observations are independent.* In other words, a subject or event is only counted once, and is not matched with or affected by another subject or event in the study.
3. *Expected frequencies cannot be too small.* There is some disagreement as to what the minimum expected frequencies can be. A conservative rule is that the minimum acceptable expected frequency for any cell is 5. If the data do not meet this requirement more data should be collected or a different statistical procedure could be used. (Turn to a more advanced text for a discussion of this topic.) Alternatively, in the case of larger chi-square designs, rows or columns could be combined in a meaningful manner so that the expected frequencies are increased.

Conclusion

We have now completed the section of the book dealing with the chi-square statistic. Before continuing with the study of additional statistical procedures it may be helpful to take a few moments to review what we have accomplished and to put it into perspective. You learned in Chapter 7 that the goodness-of-fit chi-square test is used when there is one variable and we are examining whether the observed frequencies differ from what was expected based upon the null hypothesis. We then turned in Chapter 8 to the chi-square test of independence. It is also used with frequency data but now there are two variables. In addition, we found that the null hypothesis can be stated as a difference (do the pattern of the frequencies differ) or as an association (are the variables associated).

Fortunately, much that you have learned thus far will be of use when learning the additional procedures reviewed in this book. For instance, degrees of freedom, critical values, the distinction between statistical significance and effect size, and the issue of post hoc tests will all be seen again when we review the procedures used with interval or ratio data. Accordingly, this chapter and Chapter 7 had a dual purpose. First, they served as an introduction to two statistical procedures that are employed with nominal data. Second, Chapters 7 and 8 served as a general introduction to inferential statistical procedures. It is important as you master the use of specific statistical procedures that you also learn how each new test is related to the others. It is only when you have gained this perspective that you will truly be knowledgeable of statistics.

Glossary Of Terms

Bonferroni method – A procedure to control the Type I error rate when making numerous comparisons. In this procedure the alpha level that the experimenter has set is divided by the number of comparisons.

Chi-square test of association – Another name for the chi-square test of independence.

Chi-square test of independence – An inferential procedure for analyzing whether the pattern of observed frequencies differs among the groups.

Cramer's V – Measure of effect size for chi square tests of independence larger than 2 X 2.

Dependent – Two events, samples or variables are dependent if knowing the outcome of one enhances our prediction of the other.

Effect size – A measure of how 'strong' a statistically significant outcome is.

Gambler's fallacy – The incorrect assumption that if an event has not occurred recently, then the probability of it occurring in the future increases.

p-value – The probability of an outcome, or a more extreme outcome, occurring by chance assuming the null hypothesis is correct. To be statistically significant, the p-value must be less than the alpha level, which is usually .05.

Phi – Measure of effect size for the 2 X 2 chi square tests of independence.

Post hoc comparisons – Statistical procedures utilized following an initial, overall test of *significance in order to identify the specific conditions (samples) that differ.*

References

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. 2nd Ed., Lawrence Erlbaum Associates, Hillsdale, NJ.

Mathes, E. (2003). Are sex differences in sexual vs emotional jealousy explained better by differences in sexual strategies or uncertainty of paternity. *Psychological Reports, 93*(3), 895–906.

Perfect, T. (2003). Local processing bias impairs lineup performance. *Psychological Reports, 93*(2), 393–394.

Sandson, T. A., Bachna, K. J. & Morin, M. D. (2000). Right hemisphere dysfunction in ADHD: Visual hemispatial inattention and clinical subtype. *Journal of Learning Disabilities, 33*(1), 83–90.

Questions – Chapter 8

(Answers are provided in Appendix J.)

1. When an experimenter concludes that the results of a statistical test are significant, this indicates that _____.
   a. The outcome is especially important
   b. The outcome is unlikely to have occurred by chance
   c. The experimenter has made an error

2. With a 2 X 2 chi-square, the minimum expected frequency that can occur in any cell is _____.
   a. 1
   b. 3
   c. 5
   d. 7

3. If a chi-square test is found to be significant, what measure of effect size should then be utilized?
   a. Phi
   b. Bonferroni
   c. Cramer's V
   d. Both 'a' and 'b' would always be appropriate
   e. Either 'a' or 'c' would be correct depending upon the specific chi-square

4. The chi-square test of independence is a test for _____.
   a. Difference or Association
   b. Effect size
   c. Importance
   d. None of the above

5. If there is no difference in the pattern of the observed frequencies, then a 2 X 2 chi-square will _____.

a. Be statistically significant
b. Not be statistically significant
c. Not have to meet the assumptions of the test in order to be used
d. Be more difficult to calculate

6. George, a particularly poor statistics student, notes that the last three times a coin has been tossed it has come up heads. He therefore concludes that it is time for it to come up tails. This is an example of the _____.
a. Gambler's fallacy
b. Bonferroni Method
c. Statistical significance
d. None of the above

7. Following a significant 2 X 4 chi-square test, the researcher would _____.
a. Utilize the Bonferroni Method
b. Employ phi
c. Check that all of the cells have a sample size of at least 25
d. Conduct Cramer's V
e. Both 'a' and 'd'

8. If the frequency within each cell in a 2 X 2 chi-square test is tripled, what happens to the size of the chi-square outcome?
a. It stays the same
b. It doubles
c. It triples
d. It cannot be determined

9. If the frequency within each cell in a 2 X 2 chi-square is tripled, what happens to the size of the subsequent Phi?
a. It stays the same
b. It doubles
c. It triples
d. It cannot be determined

10. The Bonferroni Method would be utilized following a statistically significant _____.
a. 2 X 2 chi-square
b. 2 X 3 chi-square
c. Phi
d. All of the above

11. With a 2 X 2 chi-square, _____ provides a measure of effect size.
a. the Bonferroni procedure
b. largest cell frequency
c. the phi–coefficient
d. none of the above

We ask freshmen and sophomores if they would like to take an arts course, and collect the following data:

	Freshmen	Sophomores
Yes	2	6
No	14	10

12. How many degrees of freedom do you have?
    a. 1
    b. 2
    c. 3
    d. 4

13. What is the value of the chi-square?
    a. 2.67
    b. 1.46
    c. 2.44
    d. 3.05

14. Does the preference for taking an arts course differ significantly between freshmen and sophomores, assuming an alpha of .05?
    a. Yes
    b. No

Now assume that the study includes juniors:

	Freshmen	Sophomores	Juniors
Yes	2	6	9
No	14	10	7

15. How many degrees of freedom do you have?
    a. 1
    b. 2
    c. 3
    d. 4

16. What is the value of the chi-square?
    a. 3.11
    b. 5.42
    c. 6.74
    d. 9.00

17. Which of the groups differs?
    a. Freshmen vs Sophomores
    b. Freshmen vs Juniors
    c. Sophomores vs Juniors
    d. Both Freshmen vs Juniors and Sophomores vs Juniors

Problems 18 – 23 utilize SPSS.

Our First Example Using SPSS With The Chi-Square Test Of Independence

To Begin SPSS

Step 1 The first step is to activate the program (Figure 5.2). Other versions of SPSS will have a very similar window. Then close the central window.

Step 2 You will see that at the bottom left of the window there are two 'switches', one labeled Data View, which is highlighted in yellow, the other Variable View (Figure 5.3).

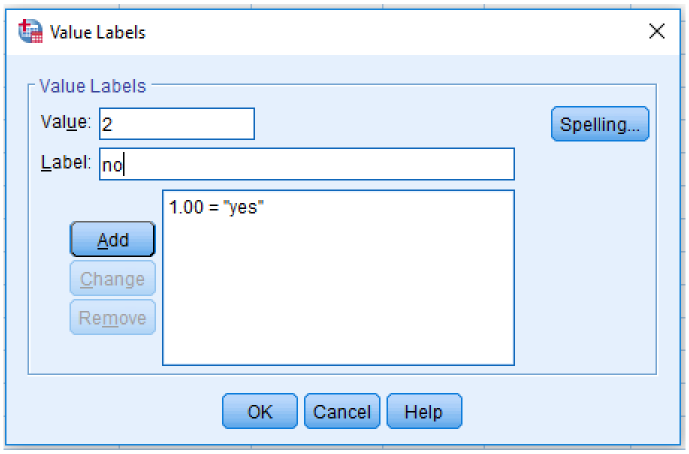
Step 3 Click on 'Variable View' at the lower left corner of the window. Near the top of the new page is a row of column headings beginning with 'Name', then 'Type', and proceeding to 'Role'. For the present we will only be dealing with the columns headed by 'Name', 'Label', 'Values' and 'Measure'.

Step 4 Click on the first empty rectangle (called a 'cell') under the column heading 'Name'. You now type the name of the first variable for which you have data. We are going to utilize the same data and labels as were previously employed in Table 8.2. These data dealt with the question of whether there is an association between whether an individual had been diagnosed with ADHD and the side of omission errors. We have called these variables 'ADHD' and 'omission'. Therefore, type 'ADHD' in the first empty cell under 'Name'.

Step 5 Click on the first empty 'cell' under the column heading 'Label'. In this cell you can type a more extensive description of your variable. In our case, type 'Diagnosed with ADHD?'. Note that in order to see the entire label you may need to expand the size of this cell by placing your cursor on the right border of the Label heading and moving to the right.

Step 6 Click on the first empty 'cell' under the column heading 'Values'. Then click on the small blue square. A box will appear. In the blank space to the right of 'Value' type the number '1'. Then type a brief description of this value of the variable in the blank space to the right of 'Label'. In our case, type 'yes'. Finally, click on 'Add'. Your label for a value of 1 will appear in the large white region in the center of the window. Now repeat the above steps in this section for the value '2', which is given the label 'no' (Figure 8.1). Click 'Add' and then click on 'OK'.

Figure 8.1 The Value Labels Window



Value Labels

Value Labels

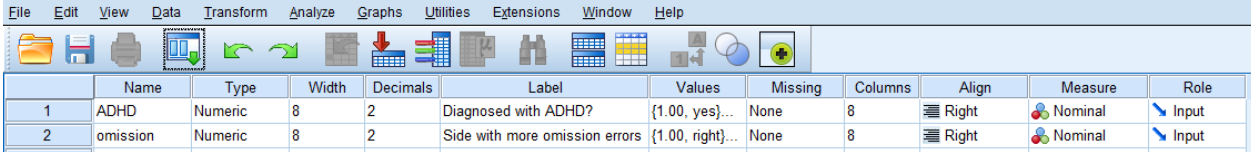
Value: 2

Label: no

1.00 = "yes"

Step 7 Click on the first empty 'cell' under the column heading 'Measure'. As we are dealing with nominal data, select 'Nominal' as is shown in Figure 8.2.

Figure 8.2 The Variable View Window



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	ADHD	Numeric	8	2	Diagnosed with ADHD?	{1.00, yes}...	None	8	Right	Nominal	Input
2	omission	Numeric	8	2	Side with more omission errors	{1.00, right}...	None	8	Right	Nominal	Input

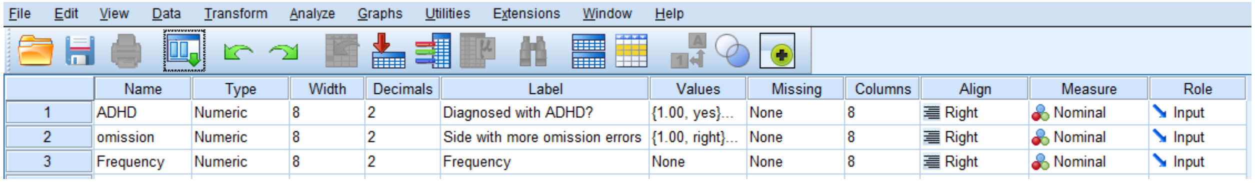
Step 8 Repeat Steps 4 – 7 except that you type 'omission' in the first empty cell under 'Name'. Then type 'Side with more omission errors' for the 'Label', and the value labels are now 'right' and 'left' instead of 'yes' and 'no'. As before, select 'Nominal' under the column heading 'Measure'. The result is shown in Figure 8.2. You could now shift to the data window and sequentially enter the data for each subject. However, this can quickly become tedious. SPSS permits the rapid construction of the chi-square data table. In order to do so we need to create another variable so that SPSS can be instructed that the numbers stand for the frequencies that occurred.

Step 9 In the empty 'cell' directly under 'omission' type the name of this new variable. I have chosen 'Frequency'.

Step 10 Move across the row and click twice on the empty 'cell' under the column heading 'Label'. In this cell you can type a more extensive description of your variable. In our case, there is no need for an extensive label, so we type 'Frequency'.

Step 11 Continue to move across the row and click on the empty 'cell' under the column heading 'Measure'. As we are dealing with frequencies, select 'Nominal', which is the SPSS designation for nominal data. This is shown in Figure 8.3. We have now completed the SPSS 'Variable View' window.

Figure 8.3 The Completed Variable View Window



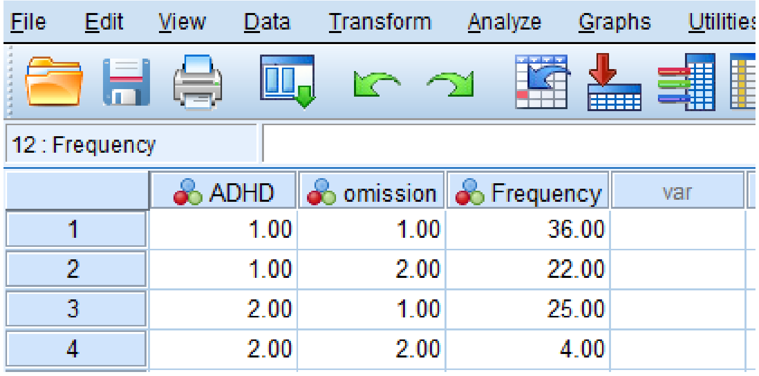
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	ADHD	Numeric	8	2	Diagnosed with ADHD?	{1.00, yes}...	None	8	Right	Nominal	Input
2	omission	Numeric	8	2	Side with more omission errors	{1.00, right}...	None	8	Right	Nominal	Input
3	Frequency	Numeric	8	2	Frequency	None	None	8	Right	Nominal	Input

To Enter Data In SPSS

Step 12 Click on the 'Data View' option at the lower left corner of the window. The variables 'ADHD', 'omission' and 'Frequency' will be present.

Step 13 Type in the values of '1' and '2' for 'ADHD' and 'omission' as shown in Figure 8.4. Each combination of these numbers specifies a chi-square cell. We need to indicate in the third column which frequency is associated with each cell of the chi-square table. The upper left cell of the chi-square table (Table 8.2) includes the data from those subjects who indicated they had been diagnosed with ADHD (condition 1 for ADHD) and who also had more omission errors on the right side (condition 1 for omission). The frequency recorded for this cell in Figure 8.4 is 36. The second cell in the first column (lower left cell of Table 8.2) includes the data from those subjects who indicated they had been diagnosed with ADHD (condition 1 of ADHD) and who also had more omission errors on the left side (condition 2 for omission). The frequency recorded for this cell in Figure 8.4 is 22. The appropriate frequencies for the remaining two cells are also indicated in Figure 8.4. It is very important that the correct frequencies are associated with each chi-square cell as is shown in Figure 8.4.

Figure 8.4 Entering Data



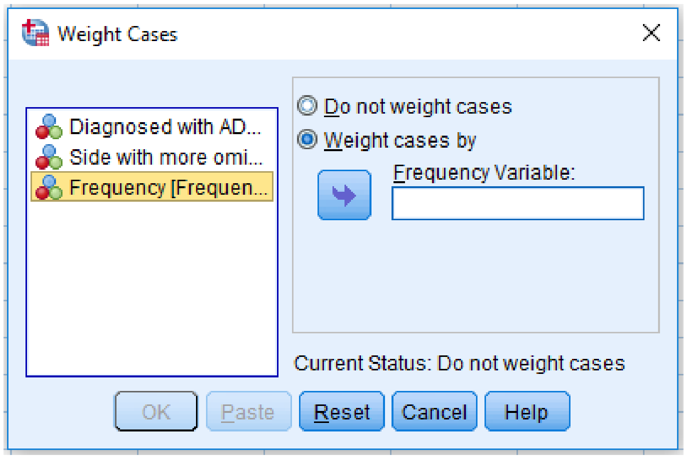
	ADHD	omission	Frequency	var
1	1.00	1.00	36.00	
2	1.00	2.00	22.00	
3	2.00	1.00	25.00	
4	2.00	2.00	4.00	

To Conduct A Chi-Square Test Of Independence

Step 14 Click your cursor on '**Data**' along the row of SPSS commands above the numbers you have entered and then move all the way down the column and click on '**Weight cases**'.

Step 15 In the new window, click on the small circle just to the left of '**Weight cases by**' and then highlight '**Frequency**' (Figure 8.5). Now click on the arrow in the center of the window. The result will look like Figure 8.6. Then click on '**OK**'. (You have just indicated to SPSS that the numbers in the variable 'Frequencies' are not scores but rather are frequencies.)
Exit the window with the statement 'WEIGHT BY frequency'. Do not save this output.

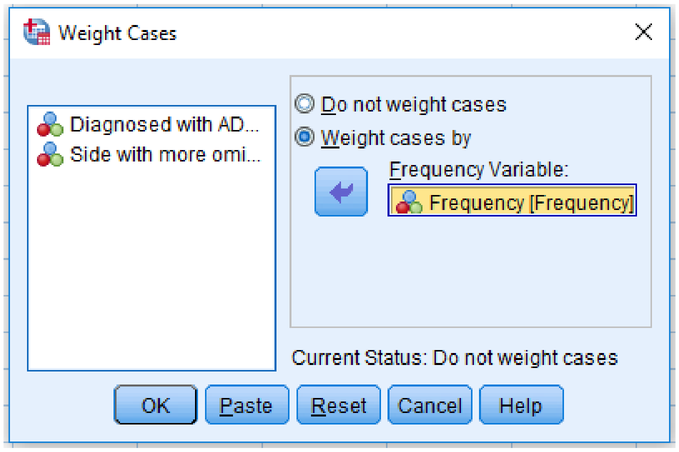
Figure 8.5 The Weight Cases Window



○ Do not weight cases
● Weight cases by
Frequency Variable:

Current Status: Do not weight cases

Figure 8.6 The Weight Cases Window



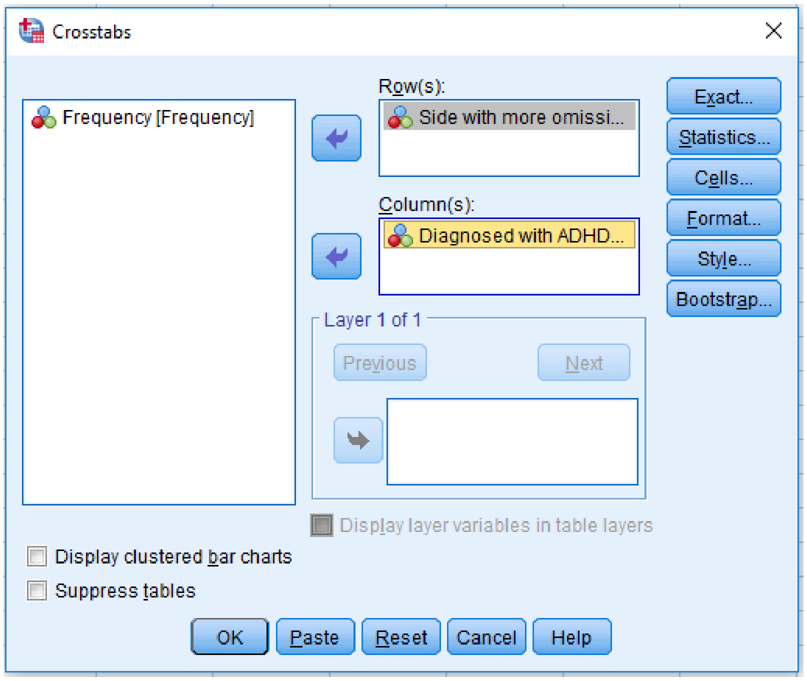
○ Do not weight cases
● Weight cases by
Frequency Variable: Frequency [Frequency]

Current Status: Do not weight cases

Step 16 Click the cursor on ‘**Analyze**’ along the row of SPSS commands above the data you entered, then move to ‘**Descriptive Statistics**’, then click on ‘**Crosstabs**’. (With SPSS you do not use the Nonparametric Statistics command with a chi-square test of independence.)

Step 17 A new window will appear. In order to recreate the rows and columns in the original data table (Table 8.2) click on ‘Side with more omission’ and then move ‘Side with more omission’ to the box under ‘Row(s)’ by clicking on the top arrow. Now move ‘Diagnosed with ADHD?’ to the box under ‘Column(s)’ by clicking on ‘Diagnosed with ADHD?’ and then clicking on the second arrow. The result will be that each label will move to the appropriate box on the right–hand side of the window, as is shown in Figure 8.7. Then click on ‘**Statistics**’ which is located in the top, right corner of the window.

Figure 8.7 Defining Crosstabs



Crosstabs

Frequency [Frequency]

Row(s): Side with more omissi...

Column(s): Diagnosed with ADHD...

Layer 1 of 1

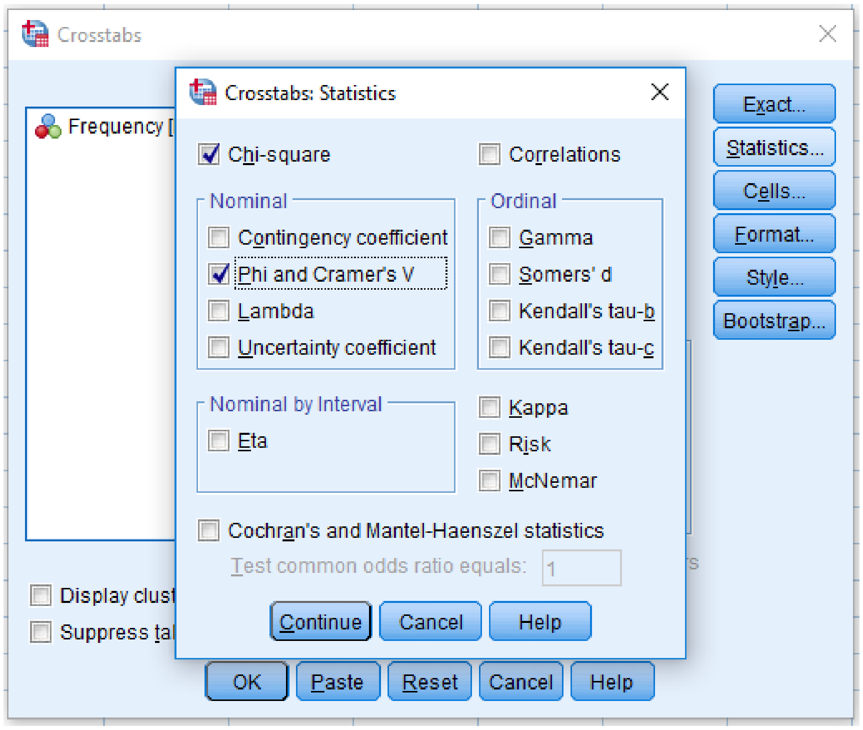
□ Display layer variables in table layers

□ Display clustered bar charts

□ Suppress tables

Step 18 A new window will appear. This window provides a number of statistical options that are available with SPSS. In this book we will limit ourselves to just a few of these, so click on the small boxes to the left of '**Chi-square**', and '**Phi and Cramer's V**' as is shown in Figure 8.8. Then click '**Continue**'.

Figure 8.8 Defining Crosstabs



Crosstabs: Statistics

☑ Chi-square

□ Correlations

Nominal

□ Contingency coefficient

☑ Phi and Cramer's V

□ Lambda

□ Uncertainty coefficient

Ordinal

□ Gamma

□ Somers' d

□ Kendall's tau-b

□ Kendall's tau-c

Nominal by Interval

□ Eta

□ Kappa

□ Risk

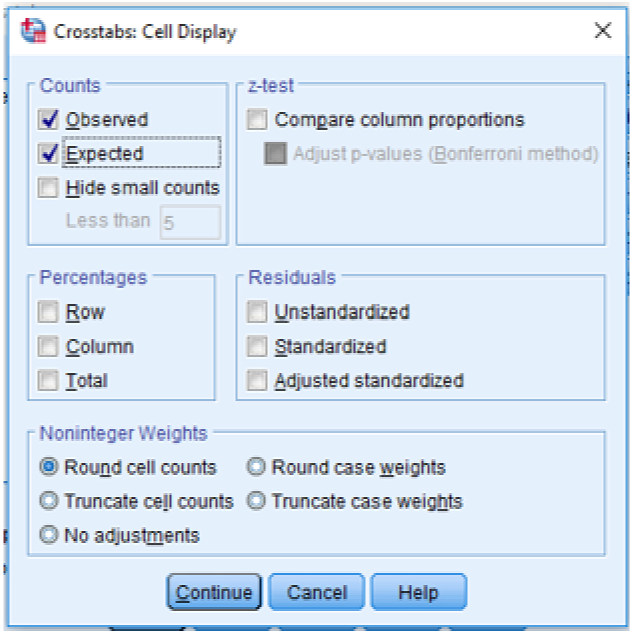
□ McNemar

□ Cochran's and Mantel-Haenszel statistics

Test common odds ratio equals: 1

Step 19 Click on ‘**Cells**’ which is located in the top, right corner of the window shown in Figure 8.8. Within the new window, be sure both ‘**Observed**’ and ‘**Expected**’ are checked, as shown in Figure 8.9. Then click on ‘**Continue**’. Now click on ‘**OK**’. SPSS provides an extensive output. We are interested in the obtained, expected and marginal frequencies (Table 8.25), the desired chi-square (Table 8.26) (SPSS calls it the Pearson Chi-square, we can ignore the other rows of the output) and finally, the effect size which is shown in Table 8.27.

Figure 8.9 Continuing to Define Crosstabs



Crosstabs: Cell Display

Counts
☑ Observed
☑ Expected
□ Hide small counts
Less than 5

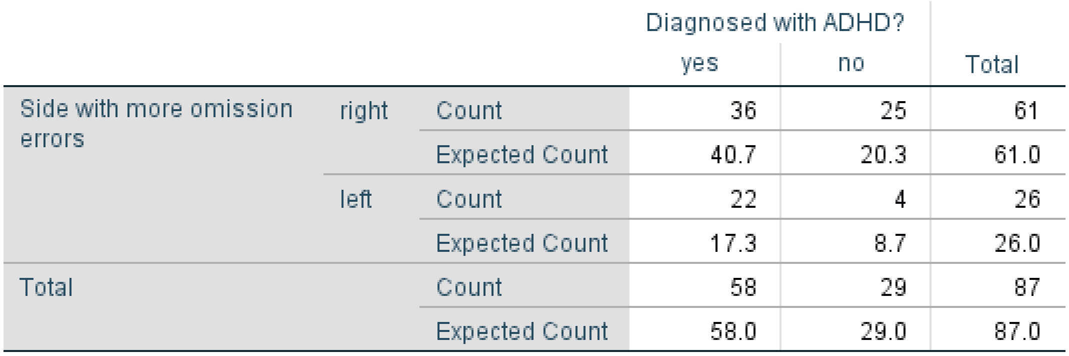
z-test
□ Compare column proportions
□ Adjust p-values (Bonferroni method)

Percentages
□ Row
□ Column
□ Total

Residuals
□ Unstandardized
□ Standardized
□ Adjusted standardized

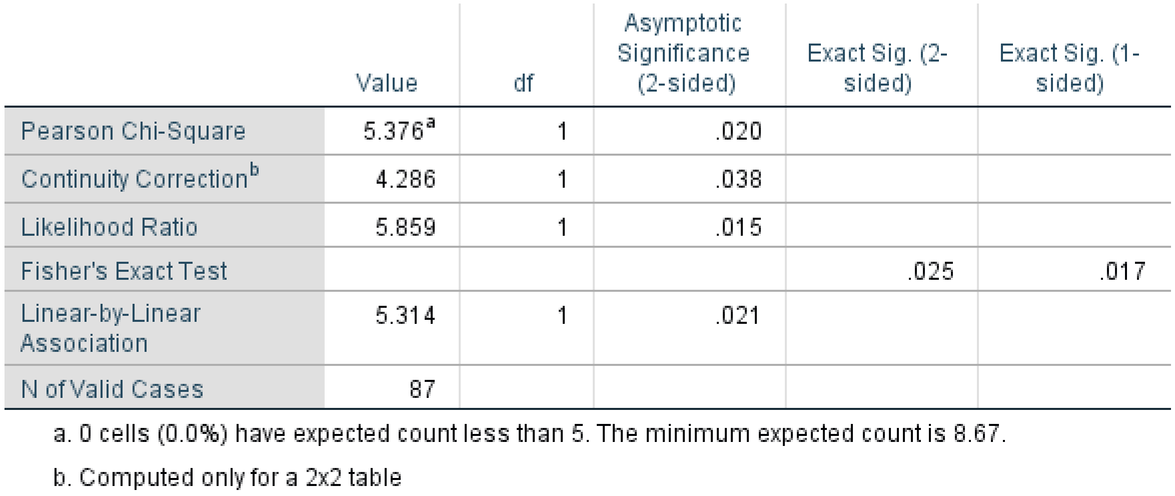
Noninteger Weights
● Round cell counts
○ Round case weights
○ Truncate cell counts
○ Truncate case weights
○ No adjustments

Table 8.25 SPSS Output; Obtained, Expected and Marginal Frequencies



			Diagnosed with ADHD?		
			yes	no	Total
Side with more omission errors	right	Count	36	25	61
		Expected Count	40.7	20.3	61.0
	left	Count	22	4	26
		Expected Count	17.3	8.7	26.0
Total		Count	58	29	87
		Expected Count	58.0	29.0	87.0

Table 8.26 SPSS Output; The Summary Chi-square Table

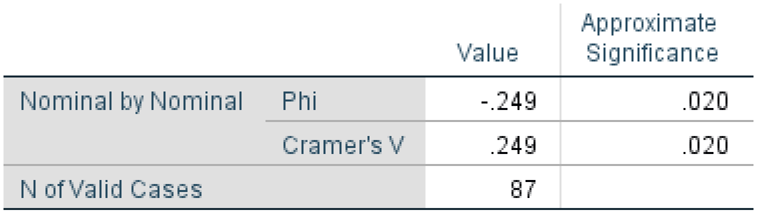


	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5.376[a]	1	.020		
Continuity Correction[b]	4.286	1	.038		
Likelihood Ratio	5.859	1	.015		
Fisher's Exact Test				.025	.017
Linear-by-Linear Association	5.314	1	.021		
N of Valid Cases	87				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 8.67.

b. Computed only for a 2x2 table

Table 8.27 SPSS Output; Effect Size



		Value	Approximate Significance
Nominal by Nominal	Phi	-.249	.020
	Cramer's V	.249	.020
N of Valid Cases		87	

You should confirm that these are the same results for the chi-square as was found earlier in this chapter except for minor rounding error when we did the calculations.

Step 20 Exit SPSS. There is no need to save your work.

Our Second Example Using SPSS With The Chi-Square Test Of Independence

To Begin SPSS

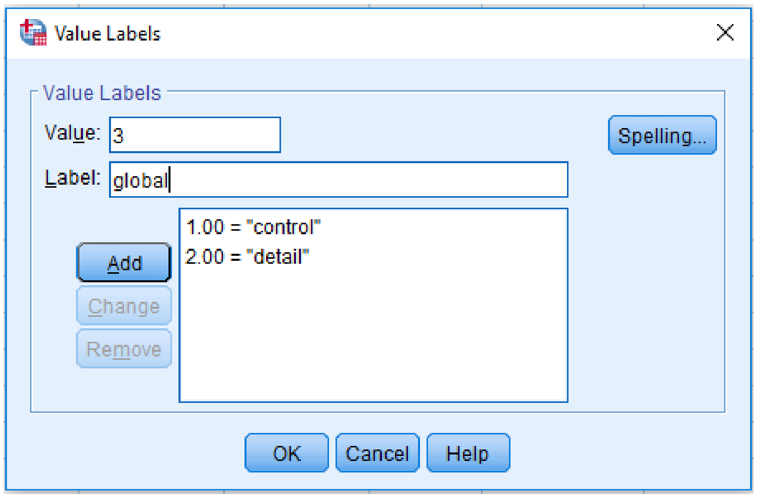
Steps 1, 2 and 3 These are the same as for the previous example.

Step 4 Click on the first empty rectangle (called a 'cell') under the column heading 'Name'. You now type the name of the first variable for which you have data. We are going to utilize the same data and labels as were previously employed in Table 8.14. These data dealt with the question of whether there is an association between the experimental condition an individual had been assigned to and their success at making a correct identification. I have called these variables 'subgroup' and 'ID'. Therefore, type 'subgroup' in the first empty cell under 'Name'.

Step 5 Click on the first empty cell under the column heading 'Label'. In this cell you can type a more extensive description of your variable. In our case, type 'Subject Group'.

Step 6 Click on the first empty ‘cell’ under the column heading ‘Values’. A box will appear. In the blank space to the right of ‘Value’, type the number ‘1’. Then type a brief description of this value of the variable in the blank space to the right of ‘Label’. In our case, type ‘control’. Finally, click on ‘Add’. Your label for a value of 1 will appear in the large white region in the center of the window. Now repeat the above steps in this section for the value ‘2’, which is given the label ‘detail’. For the value ‘3’, give the label ‘global’ (Figure 8.10). Click ‘Add’ and then click on ‘OK’.

Figure 8.10 The Value Labels Window



Value: 3

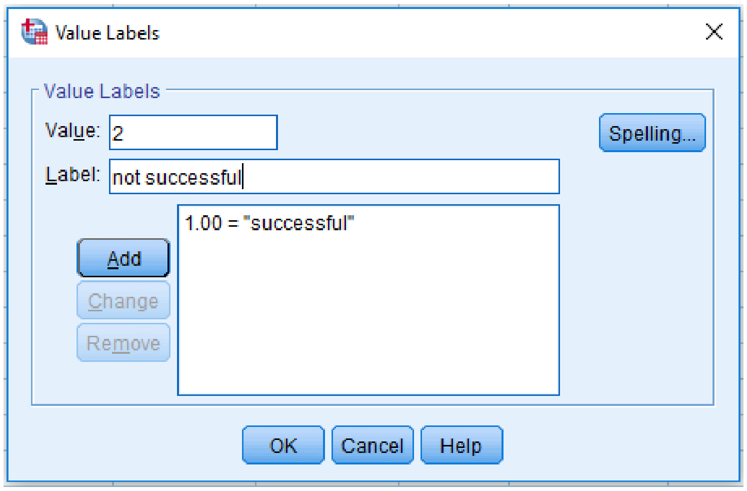
Label: global

1.00 = "control"
2.00 = "detail"

Step 7 Click on the first empty ‘cell’ under the column heading ‘Measure’. As we are dealing with nominal data, select ‘Nominal’.

Step 8 Repeat Steps 4 – 7 except that you type ‘ID’ in the first empty cell under ‘Name’, type ‘Eyewitness Success’ for the ‘Label’, and you now have two value labels, ‘successful’ and ‘not successful’, instead of ‘control’, ‘detail’ and ‘global’ (Figure 8.11). As before, select ‘Nominal’ in the column under the column heading ‘Measure’. You could now shift to the data window and sequentially enter the data for each subject. However, this can quickly become tedious. As before, we need to create another variable so that SPSS can be instructed that the numbers stand for the frequencies that occurred.

Figure 8.11 The Value Labels Window



Value Labels

Value: 2

Label: not successful

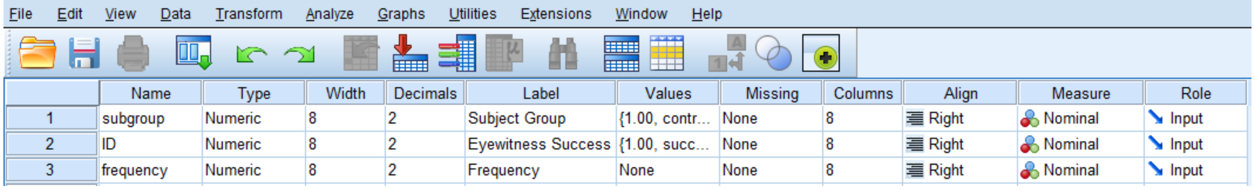
1.00 = "successful"

Step 9 In the empty 'cell' directly under 'ID' type the name of this new variable. Once again, I chose 'frequency'.

Step 10 Move across the row and click twice on the empty 'cell' under the column heading 'Label'. In this cell you can type a more extensive description of your variable. In our case there is no need for an extensive label, so we type 'Frequency'.

Step 11 Continue to move across the row and click on the empty 'cell' under the column heading 'Measure'. As we are dealing with frequencies, select 'Nominal', which is the SPSS designation for nominal data. This is shown in Figure 8.12. We have now completed the SPSS 'Variable View' window.

Figure 8.12 The Variable View Window



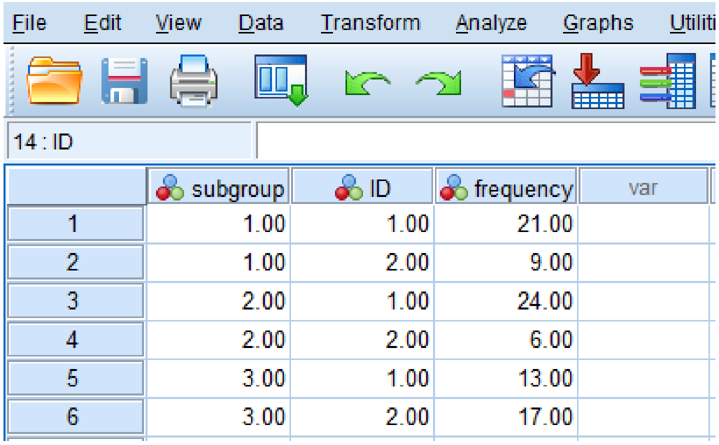
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	subgroup	Numeric	8	2	Subject Group	{1.00, contr...	None	8	Right	Nominal	Input
2	ID	Numeric	8	2	Eyewitness Success	{1.00, succ...	None	8	Right	Nominal	Input
3	frequency	Numeric	8	2	Frequency	None	None	8	Right	Nominal	Input

To Enter Data In SPSS

Step 12 Click on the 'Data View' option at the lower left corner of the window. The variables 'subgroup', 'ID' and 'frequency' will be present.

Step 13 Type in the values of '1', '2' and '3' for 'subgroup' and the values '1' and '2' for 'ID' as shown in Figure 8.13. Each combination of these numbers specifies a chi-square cell (there are 6 cells in this study) for which data in the form of a frequency are now entered in the third column. It is important that the correct frequencies are associated with each cell of Table 8.14. The result is shown in Figure 8.13.

Figure 8.13 Entering Data



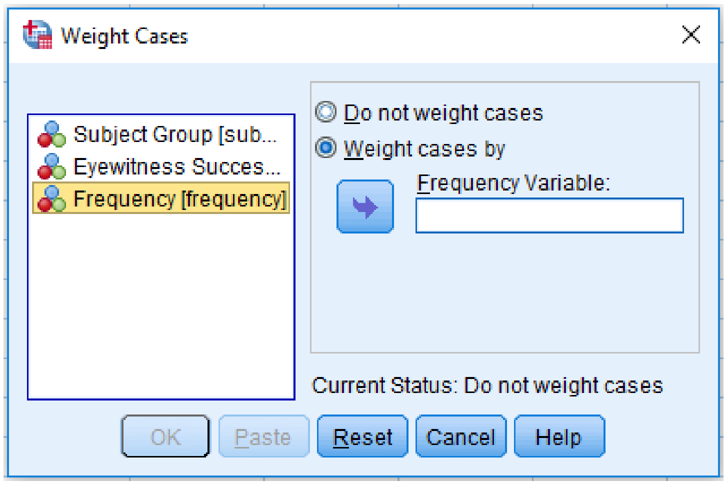
	subgroup	ID	frequency	var
1	1.00	1.00	21.00	
2	1.00	2.00	9.00	
3	2.00	1.00	24.00	
4	2.00	2.00	6.00	
5	3.00	1.00	13.00	
6	3.00	2.00	17.00	

To Conduct A Chi-Square Test Of Independence

Step 14 Click your cursor on '**Data**' along the row of SPSS commands above the numbers you have entered and then move down and click on '**Weight cases**'.

Step 15 In the new window click on the small circle just to the left of '**Weight cases by**' and then highlight '**Frequency**' (Figure 8.14). Now click on the arrow in the center of the window. The result will look like Figure 8.15. Then click on '**OK**'. (You have just indicated to SPSS that the numbers in the variable 'Frequencies' are not scores but rather are frequencies.) Exit the window with the statement 'WEIGHT BY frequency'. Do not save this output.

Figure 8.14 The Weight Cases Window



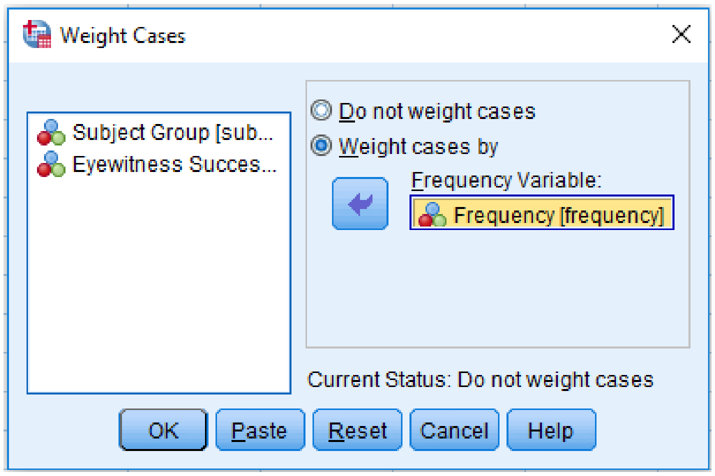
Subject Group [sub...
Eyewitness Succes...
Frequency [frequency]

○ Do not weight cases
● Weight cases by

Frequency Variable:

Current Status: Do not weight cases

Figure 8.15 The Weight Cases Window



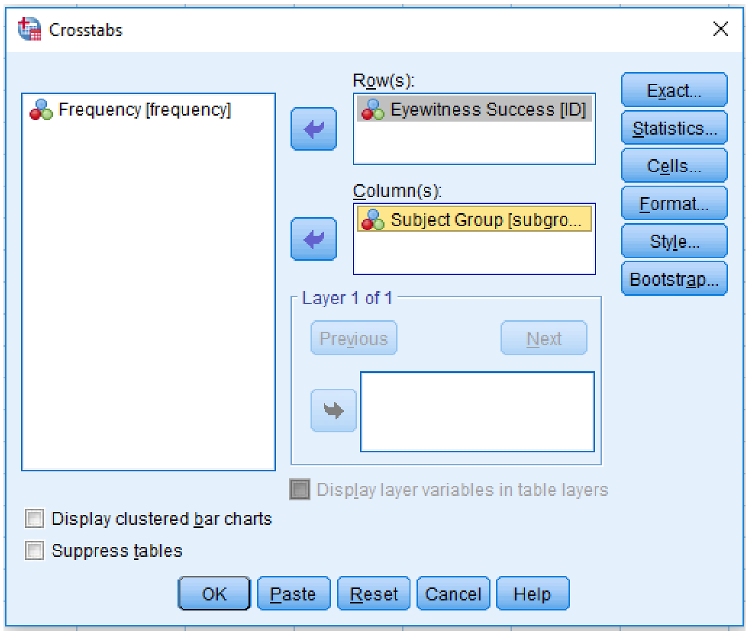
Frequency Variable: Frequency [frequency]

Current Status: Do not weight cases

Step 16 Click the cursor on '**Analyze**' along the row of SPSS commands above the data you entered, then move to '**Descriptive Statistics**', then click on '**Crosstabs**'.

Step 17 A new window will appear. In order to recreate the original rows and columns of the data (Table 8.14) move 'Eyewitness Success (ID)' to the box under 'Row(s)' by clicking on 'Eyewitness Success (ID)' and then clicking on the top arrow. Next, move 'Subject Group (subgroup)' to the box under 'Column(s)' by clicking on 'Subject Group (subgroup)' and then clicking on the second arrow. The result will be that each label will move to the appropriate box on the right–hand side of the window, as is shown in Figure 8.16. Then click on '**Statistics**' which is located in the top, right-hand corner of the window.

Figure 8.16 Defining Crosstabs

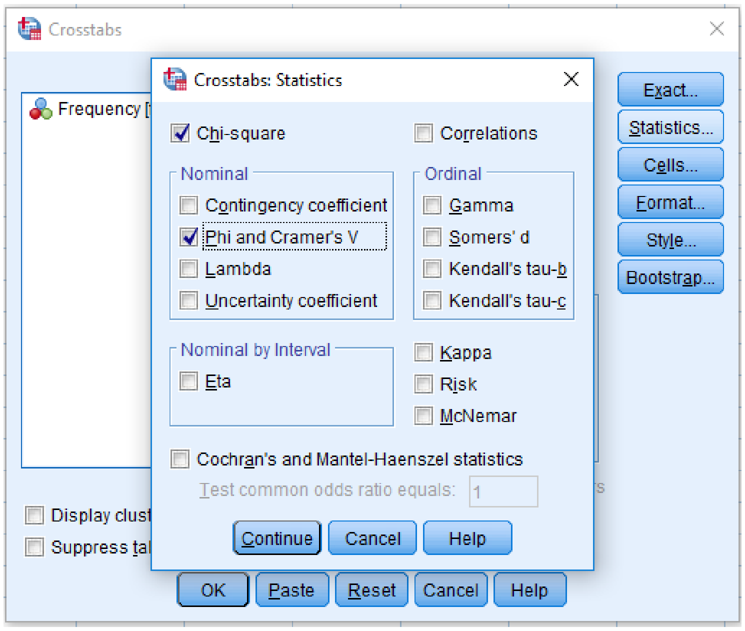


Row(s): Eyewitness Success [ID]

Column(s): Subject Group [subgro...

Step 18 A new window will appear. This window provides a number of statistical options that are available with SPSS. In this book we will limit ourselves to just a few of these, so click on the small boxes to the left of '**Chi-square**', and '**Phi and Cramer's V**' as is shown in Figure 8.17. Then click on '**Continue**'.

Figure 8.17 Defining Crosstabs



☑ Chi-square
□ Correlations

Nominal
□ Contingency coefficient
☑ Phi and Cramer's V
□ Lambda
□ Uncertainty coefficient

Ordinal
□ Gamma
□ Somers' d
□ Kendall's tau-b
□ Kendall's tau-c

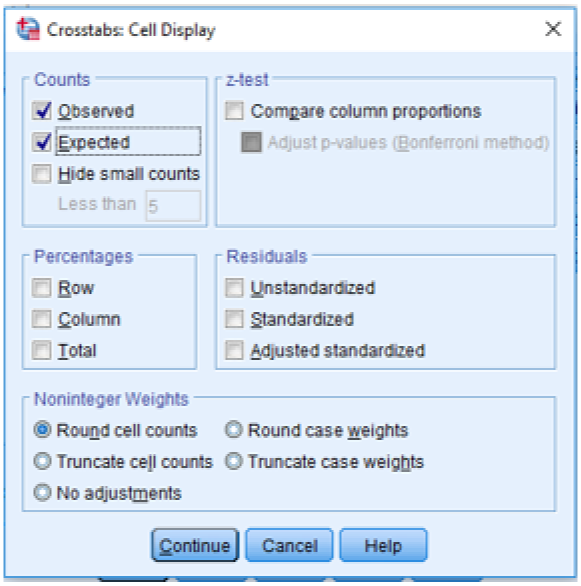
Nominal by Interval
□ Eta

□ Kappa
□ Risk
□ McNemar

□ Cochran's and Mantel-Haenszel statistics
Test common odds ratio equals: 1

Step 19 Click on '**Cells**', which is located in the top, right corner of the window shown in Figure 8.17. Within the new window click on '**Observed**' and '**Expected**' as shown in Figure 8.18. Then click on 'Continue'. Now click on '**OK**'. SPSS provides an extensive output. We are interested in the obtained, expected and marginal frequencies (Table 8.28), the desired chi-square (Table 8.29) (it is called Pearson Chi-square in SPSS, we can ignore the other rows of this output) and finally, the effect size which is shown in Table 8.30.

Figure 8.18 Continuing to Define Crosstabs



Crosstabs: Cell Display

Counts
- ☑ Observed
- ☑ Expected
- □ Hide small counts
- Less than 5

z-test
- □ Compare column proportions
- □ Adjust p-values (Bonferroni method)

Percentages
- □ Row
- □ Column
- □ Total

Residuals
- □ Unstandardized
- □ Standardized
- □ Adjusted standardized

Noninteger Weights
- ● Round cell counts
- ○ Round case weights
- ○ Truncate cell counts
- ○ Truncate case weights
- ○ No adjustments

Table 8.28 SPSS Output; Crosstabs – Obtained, Expected and Marginal Frequencies

Eyewitness Success * Subject Group Crosstabulation

			Subject Group			
			control	detail	global	Total
Eyewitness Success	successful	Count	21	24	13	58
		Expected Count	19.3	19.3	19.3	58.0
	not successful	Count	9	6	17	32
		Expected Count	10.7	10.7	10.7	32.0
Total		Count	30	30	30	90
		Expected Count	30.0	30.0	30.0	90.0

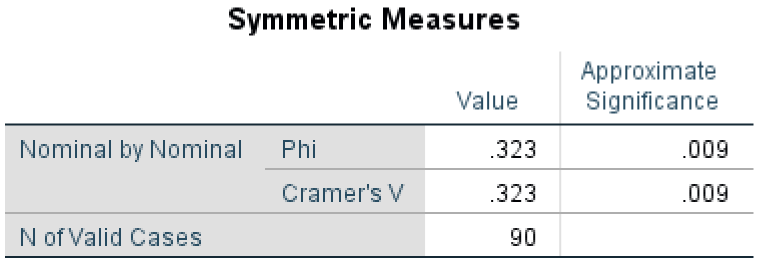
Table 8.29 SPSS Output; Summary Chi-square Table

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	9.407[a]	2	.009
Likelihood Ratio	9.417	2	.009
Linear-by-Linear Association	4.603	1	.032
N of Valid Cases	90		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 10.67.

Table 8.30 SPSS Output; Crosstabs – Effect Size



Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.323	.009
	Cramer's V	.323	.009
N of Valid Cases		90	

Once again, you should confirm that this is the same result for the chi-square as was found earlier in this chapter except for minor rounding error when we did the calculations.

Step 20 Exit SPSS. There is no need to save your work.

To confirm that you understand how to use SPSS, I suggest you redo the chi-squares that were calculated in the text for the data in Table 8.7 and Table 8.11, but this time using SPSS.

SPSS Problems – Chapter 8

Problems 18 – 23 are based upon a study by Chou, Ho and Chi (2006) which reported the frequency of depressive symptoms for Chinese older adults who were living alone or not living alone.

18. For the 90 men who lived alone, 23 reported depressive symptoms. The remainder did not report depressive symptoms. For the 851 men who did not live alone, 152 reported depressive symptoms. The remainder did not report depressive symptoms. We are interested in whether the proportions differ for these two groups of men. What is the value of the 2 X 2 chi-square?
    a. 3.183
    b. 4.791
    c. 5.482
    d. 9.236

19. How many degrees of freedom are there?
    a. 1
    b. 2
    c. 3
    d. 4

20. Is the result statistically significant if we are using a 5% region of rejection?
    a. yes
    b. no

21. For the 91 women who lived alone, 29 reported depressive symptoms. The remainder did not report depressive symptoms. For the 971 women who did not live alone, 206 reported depressive symptoms. The remainder did not report depressive symptoms. We are

interested in whether the proportions differ for these two groups of women. What is the value of the 2 X 2 chi-square?

a. 3.183
b. 4.791
c. 5.482
d. 9.236

22. How many degrees of freedom are there?

a. 1
b. 2
c. 3
d. 4

23. Is the result statistically significant if we are using a 5% region of rejection?

a. yes
b. no

Chapter 9
Finding Differences with Interval and Ratio Data – I: The One Sample z test and the One Sample t test

"Maturity is the capacity to endure uncertainty."

John Finley

Introduction

Chapter 6 introduced the scientific method. In Chapters 7 and 8 we began our discussion of inferential statistical procedures with the chi-square test, which utilizes nominal data. In the current chapter we will begin the examination of the inferential statistical procedures used with interval and ratio data (the discussion of two inferential statistical procedures used with ordinal data is given in Appendixes A and C). Our review of the inferential statistical procedures for interval and ratio data will be proceeding down the final column of Table 9.1. With interval and ratio (as well as ordinal) data you are more likely to encounter designs where there is a clear distinction between independent and dependent variables than you are when using nominal data. This difference is indicated in Table 9.1 by the separation between the columns under nominal data, and the columns under ordinal, and interval or ratio data.

We begin with the difference design that has one independent variable and only one sample of subjects. With interval or ratio data this design employs two very similar procedures, either the **one-sample z test** or the **one-sample t test**. These tests are underlined in Table 9.1. These two procedures examine the same question: Do the data collected from a single sample match what would be expected from a known or hypothesized population? Later chapters will examine designs that involve interval or ratio data and comparisons between experimental and control groups.

<u>One-sample z test</u> – An inferential procedure for comparing a sample mean with a population mean when the population standard deviation is known.

<u>One-sample t test</u> – An inferential procedure for comparing a sample mean with a population mean when the population standard deviation is not known.

Table 9.1 Overview of Inferential Statistical Procedures For Finding if there is a Difference

Type of Data		
Nominal	Ordinal	Interval/Ratio
(Frequency)	(Ranked)	(Continuous

				Measure)
Research Design		Research Design		
One Variable With At Least Two Outcomes	Goodness-of-fit Chi-Square	One IV With One Sample		One-sample z Test or One-sample t Test
		One IV With Two Or More Independent Samples	*Kruskal–Wallis H*	One-way Between–Subjects ANOVA (Only two independent samples, Independent Samples t Test)
		One IV With One Sample Having Two Or More Repeated Measures		One-way Within–Subjects ANOVA (Only two repeated measures, Dependent Samples t Test)
Two Variables, Each With At Least Two Outcomes	Chi-Square Test of Independence	Two IV Each With Two Or More Independent Samples		Two-way Between–Subjects ANOVA

The italicized procedure is reviewed in Appendix A.

The first procedure that we will review is the one-sample z test. Then we will turn to a closely related procedure, the one-sample t test.

One-Sample z Test

You are already familiar from Chapters 3 and 4 with the use of a *z score* as a descriptive statistic, and much of what you learned will be applicable here. As you recall, a z score is simply the number of standard deviations (SDs) a datum is from its mean. Furthermore, if the distribution of scores is normal then the probability of outcomes can be determined. For example, the intelligence quotient (IQ) is normally distributed and the most commonly used IQ tests have a μ of 100 and a SD of 15. An IQ score of 130 is thus 2 SDs greater than the mean, and the z score corresponding to an IQ of 130 has a value of +2 [remember, $z = (X - \mu) / \sigma$]. Use of the z table enables you to determine that approximately 98% of individuals will have an IQ less than 130. Thus 98 individuals out of 100 would be expected to have IQ scores less than 130, and only 2 out of 100 people would be expected to have IQ scores greater than 130.

We now turn to the one-sample *z test*. This is an inferential procedure. It requires that a sample is drawn randomly from a normally distributed population, and then we examine whether the obtained sample mean differs from a known or hypothetical value. For instance, let us assume that we are interested in whether engaging in a series of mental exercises will change IQ scores.

The null hypothesis would be that mental exercises will not change IQ scores (the mean of the sample would *not* differ significantly from the population mean which is 100). The alternative hypothesis would be that mental exercises will change IQ scores (the sample mean will differ from the population mean of 100).

This may appear to be just another example of the procedures that you learned for calculating a z score in Chapters 3 and 4, combined with some hypothesis testing concepts from Chapter 6. To a certain extent this is correct. However, our current use of the *z test* differs from our previous use of the *z score* in a critical way, though at first it may seem minor. The critical difference is that in our current example we are dealing with the *mean of a sample of scores* whereas Chapters 3 and 4 dealt with a *single score*.

In Chapters 3 and 4 you learned that a raw score can be converted into a z score with the following equation:

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

What this equation accomplishes is to take the difference between a score and its population mean and then divide this difference by the population standard deviation. The result is that a deviation in the original units of measurement is converted into a deviation in standard deviation units. This, in turn, permits us to determine probabilities using the z table (Appendix K, Tables 1a and 1b), assuming, of course, that the original population was normally distributed.

The equation can be rewritten as:

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

Nothing has actually changed. We have just substituted σ_X for σ. These symbols are equivalent, but with σ_X it is evident that we are referring to the variability of *scores*.

In our example of the IQs of a *sample* of subjects who have engaged in mental exercises we are not dealing with a single score. Instead we are dealing with the mean of a sample of scores. However, the logic of converting a deviation in the original units of measurement into a deviation in standard deviation units, and then referring to the z table, remains the same. The equation is simply modified to reflect that we are now comparing a sample mean (M), instead of a single score (X), to the population mean (μ). This requires, in turn, that we divide this difference by a measure of how much sample means are expected to vary. The result is as follows:

$$z = \frac{M - \mu}{\sigma_M}$$

We have a new symbol, σ_M, in this equation. Its definition can be stated in a number of ways. Perhaps the easiest definition to grasp is that it is a measure of how much *sample means* are expected to vary. More precisely, though probably not as clearly, it is the standard deviation of the means of samples of a given size selected from a single population. Alternatively, it can be defined as the standard deviation of the sampling distribution of means, which is equivalent to saying it is the standard deviation of the population of sample means. It is encountered so frequently that it has its own name, the **standard error of the mean (SEM)**.

Standard error of the mean (SEM) – The standard deviation of the sampling distribution of means.

It is important that you recognize the parallel between obtaining a z score, which was described in Chapters 3 and 4, and conducting a z test:

To find a z score

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

To conduct a z test

$$z = \frac{M - \mu}{\sigma_M}$$

In each case we are converting a deviation, either $(X - \mu)$ or $(M - \mu)$, into standard deviation units. This requires that we divide the obtained deviation by the appropriate measure for the standard deviation. In the case of $X - \mu$ we are dealing with how much a *score* deviates from its population mean and so we divide by the standard deviation of scores (σ_X). In the case of $M - \mu$ we are dealing with how much a *sample mean* deviates from its population mean and so we divide by the standard deviation of sample means, which is also called the standard error of the mean (SEM) which also has the symbol σ_M. In either case we then refer to the z table. However, before actually conducting a one-sample z test we need to first gain a better understanding of the standard error of the mean (σ_M).

The Standard Error Of The Mean (SEM)

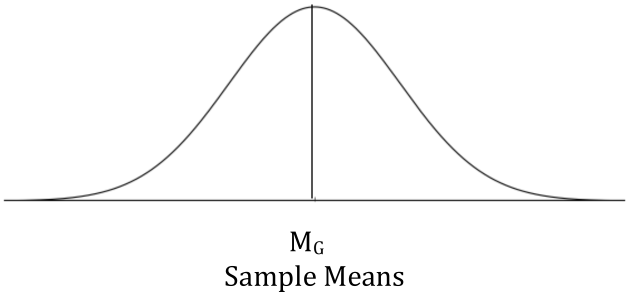
In our current example, which is examining the effect of mental exercises upon IQ, we began by randomly selecting a sample. We would not be surprised if the mean IQ score of our sample differed slightly from the population mean of 100 even before experiencing the mental exercises. Though the sample was randomly selected, some variation would be expected. This discrepancy between a population's mean and the mean of a sample drawn from it was not caused by any action of the experimenter. Instead it was due to chance. And in statistics we would identify this difference as being an example of **error**. This does not signify that someone made a mistake. It simply indicates that the outcome is due to chance events.

Error – An outcome due to chance.

Since we have shown that it is not surprising if a sample's mean differed from the population mean, it follows that if we selected a large number of samples, all of the same size and from the same population, we would also expect to find variability among their means and the population mean. But how much variability? While small variations would be expected, large variations would be less likely. The reason for this is that in order to obtain a large difference a sample would have to have a preponderance of subjects with either very low or very high IQs. Since the samples are being selected randomly, this is not likely to occur. Consequently, if we graphed the means of a large number of samples, all of the same size, we would expect to obtain a distribution such as in Figure 9.1. This distribution is called the **sampling distribution of the mean**.

<u>Sampling distribution of the mean</u> – A theoretical probability distribution of sample means. The samples are all of the same size and are randomly selected from the same population.

Figure 9.1 The Sampling Distribution of the Mean



Notice again that the sampling distribution of the mean is a graph of *sample means* (M). Thus each point in the distribution is an M. If we wanted to find the 'average' of the sample means we could, of course, find the mean of these sample means. This probably sounds a bit strange, but the mean of these sample means, or the **grand mean**, which is represented in Figure 9.1 by the symbol M_G, provides an excellent estimate of the population mean which, you recall, has the symbol μ. Furthermore, it can be shown mathematically that if the samples are being selected from a normally distributed population, such as IQ, then the sampling distribution of the mean is also normally distributed. Thus it is symmetrical as well as bell-shaped, with most of the sample means grouped near the middle of the distribution (Figure 9.1).

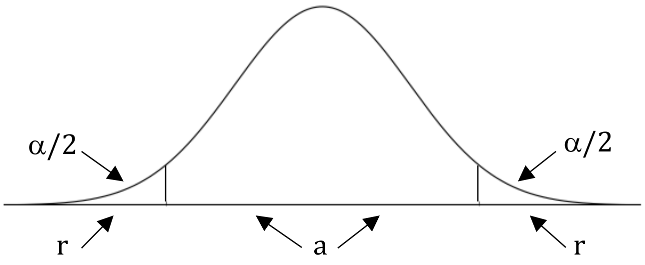
<u>Grand mean</u> (M_G) – The mean of the sample means.

One- And Two-Tailed Tests

We have just noted that sample means are expected to vary from the population mean and from each other, even without the experimenter introducing a treatment. This raises a question, how discrepant do the sample and population means have to be after a treatment is introduced in order to reject the null hypothesis that an observed difference is simply due to chance variation? In other words, using our example, how large a difference must be observed to conclude that the mental exercises had an effect on the IQ scores? There is no absolute answer to this question. However, if the difference between the sample and population means is due to chance then it is expected that in most cases this difference will be small. In only a few cases would a large discrepancy be expected to happen just by chance. Thus, the larger the difference between the sample and population means, the less likely this difference is due to chance, and the more likely it is due to the effect of the treatment.

As was noted in Chapter 6, in most fields it has come to be accepted that if an outcome would be expected to occur, by chance, less than 1 time in 20 we reject the null hypothesis and accept the alternative hypothesis. One time in 20 is equivalent to .05 or 5%. It was also pointed out that there is nothing magical about .05. A different criterion such as .01 can be, and sometimes is, chosen. If a criterion of .01 is chosen, then we retain the null hypothesis unless an outcome is so unlikely that it would be expected to occur in less than 1 out of 100 cases by chance. The criterion chosen is, of course, the alpha level and its symbol is the Greek letter, α. Recall from Chapter 7 that the critical region encompasses the most extreme possible outcomes. In the current example the critical region would be divided into two portions of the sampling distribution of the mean. This is because our null hypothesis does not state whether the treatment (engaging in a series of mental exercises) is expected to increase or decrease the IQ scores of the sample. Consequently, as our total area of rejection is equal to α, the probability of each of the two critical regions is equal to α / 2 (Figure 9.2).

Figure 9.2 Two-Tailed Test



In the current example, the null hypothesis was that the treatment (mental exercises) would not have an effect. The alternative hypothesis was that the treatment would have an effect. The null hypothesis could be rejected, therefore, if the sample's mean IQ was either much lower or much higher than the population's mean IQ. Because an extreme outcome in either direction would result in the rejection of the null hypothesis, this is called a **two-tailed or nondirectional test**. In a few moments we will see that a directional, or one-tailed test, is also possible. For now, it is important you understand that with a two-tailed test, if the mean of the sample is so different from the population mean that it falls in the region or area of rejection on the sampling distribution of the mean, indicated by the letter 'r' in each tail of Figure 9.2, then the null hypothesis will be rejected. And if we reject the null hypothesis then the alternative hypothesis will be accepted. Of course, the sample mean might not differ so much from the population mean that it falls in the extreme tails represented by the letter 'r'. In this case it would fall in region 'a' of Figure 9.2 and the null hypothesis would not be rejected. Instead the null hypothesis would be accepted or retained.

<u>Two-tailed or nondirectional test</u> – An analysis in which the null hypothesis will be rejected if an extreme outcome occurs in either direction. In such a test, the area of rejection is divided into two parts, each equal to α / 2.

If the alternative hypothesis in our IQ study had been that the mental exercises would *increase* the mean IQ of the sample, we would now have a directional prediction. In this case, the entire region of rejection would be put in the upper tail of the sampling distribution of the mean. This is illustrated in Figure 9.3. Similarly, if the original alternative hypothesis had been that the mental exercises would *decrease* the mean IQ of the sample, this would also be a directional prediction. In this case, however, the entire region of rejection would be put in the lower tail of the distribution, as is illustrated in Figure 9.4. As will be shown shortly, putting the entire region of rejection in one tail has the advantage that a smaller difference between the sample and population means is needed in order for us to reject the null hypothesis. However, in order to reject the null hypothesis the result must be in the predicted direction. With a directional prediction, no matter how large the observed difference, if it is in the direction opposite to what was predicted the null hypothesis is retained. In other words, if you use a **one-tailed or directional test** you are 'putting all of your eggs in one basket'. For this reason, one-tailed tests are *much less commonly used* than two-tailed tests. It is also important to note that the decision whether you have a directional or non-directional hypothesis is based upon the results of previous research and must be made before you collect any data. It would be unethical to collect your data, determine the direction of the difference and then decide to use a one-tailed test.

<u>One-tailed or directional test</u> – An analysis in which the null hypothesis will only be rejected if an extreme outcome occurs in the predicted direction. In such a test, the single area of rejection is equal to alpha and it is located in one tail of the sampling distribution.

Figure 9.3 One-Tailed Test with the Area Of Rejection in the Upper Tail

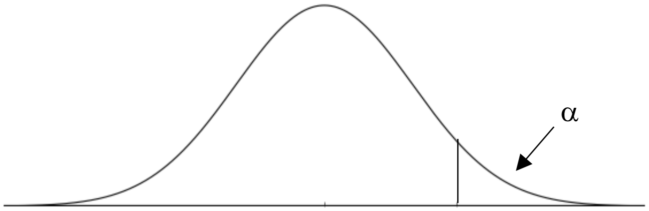
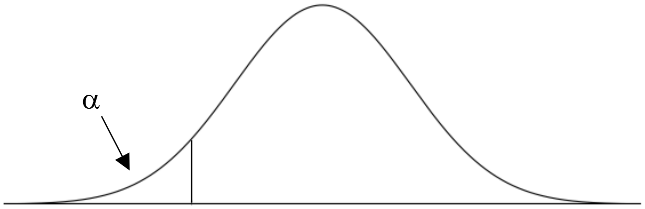


Figure 9.4 One-Tailed Test with the Area Of Rejection in the Lower Tail



As was just noted, if you have enough information from previous research to choose a one-tailed test this will make it more likely you will reject the null hypothesis compared to using a two-tailed test, and thus will increase the statistical power. Of course, this assumes that your results actually turn out to be in the predicted direction. If they don't, your only recourse is to conduct the entire study over again, this time using a two-tailed test. Remember, you have put 'all of your eggs in one basket' with a one-tailed test.

Summarizing to this point, when conducting an experiment we tentatively accept that the null hypothesis is true unless there is sufficient evidence from the experiment to indicate that this is unlikely. The criterion for deciding how unlikely the outcome must be in order to reject the null hypothesis is set by the experimenter when the alpha level is chosen. If the null hypothesis is rejected, we then tentatively accept that the alternative hypothesis is correct. Remember, with statistics we have not 'proven' that the alternative hypothesis is correct. And though we are making informed decisions we recognize that making an error is still possible. However, the use of the statistical procedures outlined in this text will reduce the likelihood of making an incorrect decision.

To determine whether there is sufficient evidence to reject the null hypothesis we need to locate the position of our sample's mean on the theoretical frequency distribution of all possible sample means which, you recall, is called the sampling distribution of the mean. This requires that we have a measure of the variability of this distribution.

It should be obvious that the sample means will differ from each other less than the individual scores will differ in the population from which these samples are drawn. For instance, it is possible that a single score randomly selected from a population will be extreme, and thus differ substantially from the population mean; but it is unlikely that even a relatively small sample would consist entirely of extreme scores and all of them in the same direction. As a result, the means of samples will be grouped more closely around the population mean than the original scores were. Furthermore, the larger the sample, the less likely it is that it would consist entirely of an extreme group of subjects. Thus, the larger the randomly selected samples, the closer their means are expected to be to the population mean. This is an example of what is known in the field of statistics as the **law of large numbers**. It states that the larger the size of the random sample the better the estimate of population parameters such as the mean. How much the variability of sample means (σ_M) is reduced as the sample size increases is determined with the following equation:

$$\sigma_M = \frac{\sigma_X}{\sqrt{n}}$$

The equation states that σ_M, the standard error of the mean, which is the standard deviation of *sample means,* is equal to σ_X, the standard deviation of *scores*, divided by $\sqrt{}$n, the square root of the sample size. Thus, as the sample size increases, σ_M will decrease, just as we reasoned.

Law of large numbers – The larger the sample size, the better the estimate of population parameters such as μ.

The equation $\sigma_M = \sigma_X/\sqrt{n}$ indicates that there is a relationship between the variability of the sample means (σ_M), the variability of the scores (σ_X), and the sample size (n). Specifically, the variability of sample means, also called the standard error (σ_M), will *increase* as the variability of the scores (σ_X) *increases* or the sample size (n) *decreases*. Alternatively, the variability of sample means (σ_M) will *decrease* as the variability of the scores (σ_X) *decreases* or the sample size (n) *increases*. These relationships will be clearer with some examples. We will begin by varying the sample size, n.

Let us assume that our sample consisted of 4 subjects. How will σ_M and σ_X be related? Remember:

$$\sigma_M = \frac{\sigma_X}{\sqrt{n}}$$

With a sample size of 4 this becomes:

$$\sigma_M = \frac{\sigma_X}{\sqrt{4}}$$

$$= \frac{\sigma_X}{2}$$

Thus, if the sample size is 4, σ_M (the variability of sample means) will be equal to only 1/2 of σ_X (the variability of scores). Put another way, if the sample size is 4, the means of these samples are expected to vary only half as much as the scores vary.

What if our sample consisted of 9 subjects? Now how will σ_M and σ_X be related?

$$\sigma_M = \frac{\sigma_X}{\sqrt{n}}$$

$$= \frac{\sigma_X}{\sqrt{9}}$$

$$= \frac{\sigma_X}{3}$$

Thus, if the sample size is 9, σ_M (the variability of sample means) will be equal to only 1/3 of σ_X (the variability of scores). Put another way, if the sample size is 9 the means of these samples are expected to vary only one third as much as the scores vary.

What if the sample size was increased to 25?

$$\sigma_M = \frac{\sigma_X}{\sqrt{n}}$$

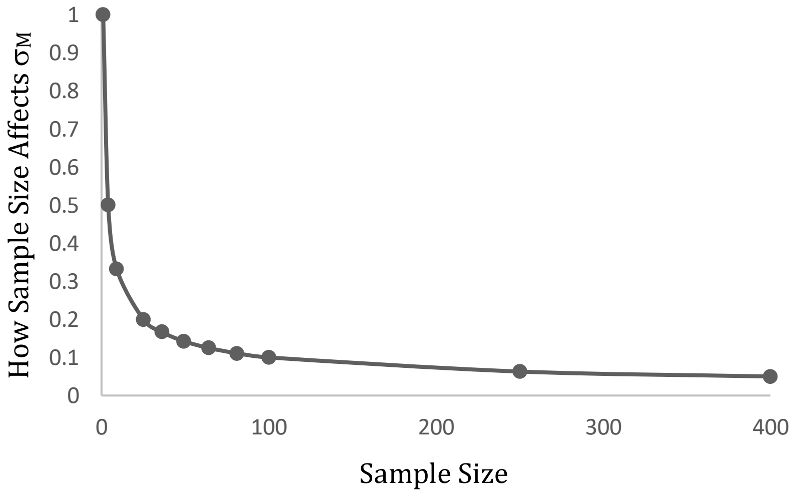
$$= \frac{\sigma_X}{\sqrt{25}}$$

$$= \frac{\sigma_X}{5}$$

Thus, if the sample size is 25, σ_M will be reduced to only 1/5 of σ_X.

The relationship between the sample size and the magnitude of the standard error is shown in Figure 9.5. Clearly, as the sample size increases, the variability of σ_M is decreasing. And remember, σ_M is due to chance events, what we call error. It is not due to our treatment. So, increasing the sample size will decrease the error that is expected. But, most of the decrease in the error occurs by the time you get to a sample size of 25 or 30. And increasing the sample size beyond 100 has almost no effect. In other words there is virtually no improvement in how well a sample mean matches the population mean with samples greater than 100.

Figure 9.5 Relationship Between the Standard Deviation, Standard Error and the Sample Size



For example:

With a sample size of 4, $\sigma_M = .50\ \sigma_X$

With a sample size of 9, $\sigma_M = .33\ \sigma_X$

With a sample size of 100 $\sigma_M = .10\ \sigma_X$

With a sample size of 250, $\sigma_M = .063\ \sigma_X$

With a sample size of 400, $\sigma_M = .05\ \sigma_X$

Now we will explore what happens to the magnitude of the standard error when the variability of the scores, in other words the value of σ_X, is increased. Specifically, if σ_X is doubled from 15 to 30 while the sample size remains constant with an n of 9, how does σ_M change?

When $\sigma_X = 15$ and $n = 9$:

$$\sigma_M = \frac{\sigma_X}{\sqrt{n}}$$
$$= \frac{15}{\sqrt{9}}$$
$$= \frac{15}{3}$$
$$= 5$$

And when $\sigma_X = 30$ and $n = 9$:

$$\sigma_M = \frac{\sigma_X}{\sqrt{n}}$$
$$= \frac{30}{\sqrt{9}}$$
$$= \frac{30}{3}$$
$$= 10$$

Thus, as the variability of the scores doubles, in other words as the value of σ_X doubles, so does the magnitude of σ_M. Put another way, samples drawn from more variable populations are expected to vary more than samples drawn from less variable populations.

As was previously noted, it can be proven mathematically that if you randomly select a large number of samples, all of the same size, from a normally distributed population, the distribution of these sample means (the sampling distribution of means) will also be normally distributed. It is unlikely that you would ever do this, but the conclusion is important. Further, you have learned that the mean of these sample means (M_G) would be an excellent predictor of the population mean (μ). And the standard deviation of the sampling distribution of means (the standard error, σ_M) will equal the standard deviation of population scores (σ_X) divided by the square root of the sample size (n). These critically important conclusions are summarized in what is known as the **central limit theorem**.

The Central Limit Theorem

Many inferential procedures assume that a sample is being drawn from a normally distributed population. This assumption is necessary because these tests are based upon sampling distributions. In the case of the one-sample z test this is the sampling distribution of the mean. And we have noted that if the underlying population of scores from which the sample is drawn is normally distributed, then the sampling distribution of the mean will also be normal. It may have occurred to you, however, that this raises a problem, for how would a researcher possibly know if the population they are interested in actually has a normal distribution unless, as with the IQ test and SAT, this has been previously determined?

Fortunately, it can be shown mathematically that as the sample size increases, the shape of the distribution of sample means (the sampling distribution of the mean) rapidly approximates the normal distribution *irrespective of the shape of the population from which it is drawn*. If the population is normally distributed, then the shape of the distribution of sample means will be normal regardless of the sample size. Furthermore, if the population closely approximates being normal, then even with a small sample size the shape of the distribution of sample means will be close to normal. However, if the population is markedly non-normal, then the sample size will have to be larger before the shape of the distribution of sample means will approach being normal. Unfortunately, since we often don't know what the shape of the original population is, we don't know precisely how large our sample needs to be in order to result in a sampling distribution of the mean that is approximately normally distributed. As a general rule of thumb, so long as the sample has 30 or more subjects you can safely assume the sampling distribution is essentially normal. In addition, according to the central limit theorem, the mean of the sample means making up a sampling distribution (M_G) is an excellent predictor of the population mean (μ). Finally, the standard deviation of the sampling distribution (σ_M) is equal to $\sigma_X/\sqrt{n}$.

Central limit theorem –

–With increasing sample sizes, the shape of the distribution of sample means (sampling distribution of the mean) rapidly approximates the normal distribution irrespective of the shape of the population from which it is drawn.

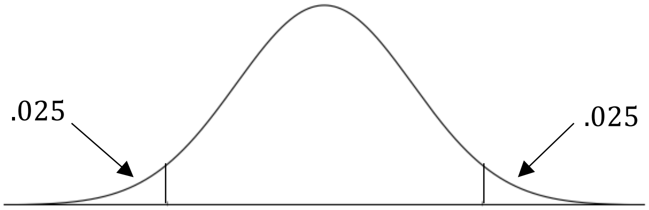
–The mean of the distribution of sample means (M_G) is an unbiased estimator of the population mean.

–And the standard deviation of the distribution of sample means (σ_M) will equal $\sigma_X/\sqrt{n}$.

Conducting A One-Sample z Test

Returning to our example of whether mental exercises affect IQ, let us assume that, based upon a sample size of 25, the mean of the sample of IQ scores was 105. Recall that IQ scores are essentially normally distributed, the mean of the population of IQ scores is known to be 100, and the standard deviation of IQ scores is 15. The null hypothesis would state that any difference between the sample and population means is due to chance. The alternative hypothesis is that this observed difference is indicative of a true difference existing between the sample and population means. The one-sample z test can be used to decide between these two hypotheses. As is common we set alpha equal to .05. Since no direction for the outcome has been predicted, this is a two-tailed test and thus one half of the area of rejection will be located in each tail of the theoretical frequency distribution (the sampling distribution of the mean). In other words, the area of rejection will consist of .05 / 2, which equals .025, in each tail (Figure 9.6). From the z table we ascertain that an area of .025 in the lower tail of the distribution is equivalent to a z of −1.96 (Appendix K, Table 1a), and the area of .025 (which is equivalent to a percentile rank of 0.975) in the upper tail is equivalent to a z of +1.96 (Appendix K, Table 1b). In order to reject the null hypothesis our outcome would need to be more extreme than one of these z scores.

Figure 9.6 The Two-tailed Test



We can now conduct our z test:

$$z = \frac{\mathbf{M - \mu}}{\sigma_{\mathbf{M}}}$$

M and μ are known and can be substituted directly into the equation:

$$z = \frac{\mathbf{105 - 100}}{\sigma_{\mathbf{M}}}$$

However, σ_M needs to be calculated from the equation:

$$\sigma_M = \frac{\sigma_{\mathbf{X}}}{\sqrt{\mathbf{n}}}$$

$$= \frac{\mathbf{15}}{\sqrt{\mathbf{25}}}$$

$$= \frac{\mathbf{15}}{\mathbf{5}}$$

$$= 3$$

We can now substitute this value into our equation for z:

$$z = \frac{\mathbf{105 - 100}}{\sigma_{\mathbf{M}}}$$

$$= \frac{\mathbf{105 - 100}}{\mathbf{3}}$$

$$= \frac{\mathbf{5}}{\mathbf{3}}$$

$$= 1.67$$

Recall that we previously determined that an area of .025 in the tails of the distribution is equivalent to a critical value of + or −1.96 (Appendix K, Tables 1a and 1b). Thus an obtained z beyond +/ − 1.96 leads us to reject the null hypothesis. With an obtained value within +/− 1.96 we do not reject the null hypothesis. In our example the obtained z of 1.67 is less than the critical value of 1.96, so we retain the null hypothesis. Stated differently, our obtained sample mean of 105 is 1.67 standard units from the population mean of 100. Since we are dealing with the deviation of a sample mean from the population mean we could also say that the sample mean of 105 is 1.67 standard errors from the population mean of 100. However, to reject the null hypothesis we would need a sample mean more than 1.96 standard errors from the population mean.

Reporting The Results Of An Insignificant One-Sample z Test

In an article we would state, "There was not sufficient evidence that the IQ scores of subjects who had engaged in mental exercise ($M = 105$, $SEM = 3$) differed from the expected population value ($z = 1.67$, $p > .05$)". It is important to note the direction of the > symbol.

Examining The Effect Of Using A One-Tailed Test

It is instructive to reexamine the previous example assuming that a one-tailed test had been appropriate. Remember, choosing a one-tailed test would need to have occurred before the data were collected, but for illustration purposes let us assume that prior research had suggested that mental exercises would have increased IQ scores. If so, we would have been justified in utilizing a one-tailed test with all of the rejection area in the high end of the curve. The critical value of z, with alpha equal to .05, would be found by looking for an area of .95 in the body of the z table (Appendix K, Table 1b). The middle of the range of critical values would be +1.64 or +1.65. As this is a one-tailed test the direction of the outcome is important. The null hypothesis would now be that mental exercises would increase IQ and thus we would reject the null only if the obtained value of z was greater than the new critical value. As our obtained z of +1.67 is greater than the critical value, and is in the predicted direction, we would now have a statistically significant difference. Clearly, whether you initially choose to conduct a one- or two-tailed test can matter.

Another Example Of The One-Sample z Test

To be certain that you understand the use of the one-sample z test we will review another example. Let us assume that we are interested in ascertaining whether high school students with low grade point averages have the same SAT scores as the general population of students who took the test. The null hypothesis would be that the mean SAT of these students is the same as the mean SAT for the general population, which is 500. The alternative hypothesis would be that the mean SAT score of these students differs from what is found with the general population. As no direction has been specified for an outcome this is a two-tailed test and, as usual, we set alpha equal to .05. In order to differentiate between the null and alternative hypotheses we collect SAT scores from a random sample of 49 high school students who have low grade point averages. We find that the mean SAT of this sample is 467. As it is known that the standard deviation of the SAT test is 100, and that the SAT test is normally distributed, we can now conduct our one-sample z test:

$$z = \frac{\mathbf{M - \mu}}{\sigma_{\mathbf{M}}}$$

M and μ are known and can be substituted directly into the equation:

$$z = \frac{\mathbf{467 - 500}}{\sigma_{\mathbf{M}}}$$

As before, σ_M needs to be calculated from the equation:

$$\begin{aligned}\sigma_M &= \frac{\sigma_{\mathbf{X}}}{\sqrt{\mathbf{n}}}\\ &= \frac{\mathbf{100}}{\sqrt{\mathbf{49}}}\\ &= \frac{\mathbf{100}}{\mathbf{7}}\\ &= 14.29\end{aligned}$$

We can now substitute this value for the standard error into our equation for z:

$$\begin{aligned}z &= \frac{\mathbf{467 - 500}}{\sigma_{\mathbf{M}}}\\ &= \frac{\mathbf{467 - 500}}{\mathbf{14.29}}\\ &= \frac{\mathbf{-33}}{\mathbf{14.29}}\\ &= -2.31\end{aligned}$$

Recall that for a two-tailed test with an alpha equal to .05 the critical value is + or –1.96. As our obtained z of –2.31 is beyond (more extreme than) the critical value, we reject the null hypothesis that there is no difference between the SAT scores of high school students with low grade point averages and the SAT scores of the general population, and we accept the alternative hypothesis that there is a difference.

Reporting The Results Of A Significant One-Sample z Test

In an article, we could say, “There was evidence that high school students who had low grade point averages had lower SAT scores ($M = 467$, $SEM = 14.29$) than the general population of students who have taken the exam ($z = -2.31$, $p < .05$)”.

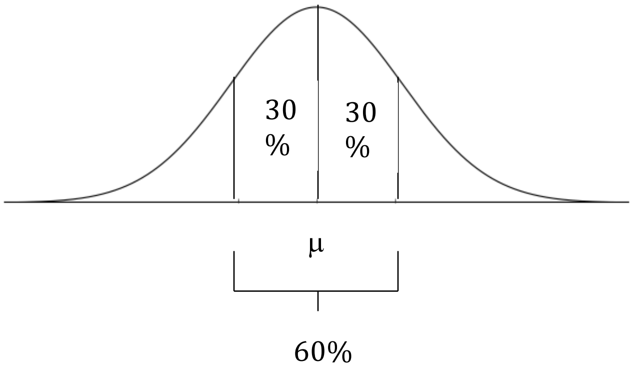
Finding A Confidence Interval For z

The one-sample z test deals with whether the difference between the sample and population means is sufficient to reject the null hypothesis. Alternatively, we could take the data from our sample and instead ask what is the range of values that has a known probability of including the population mean. For instance, let us assume that a sample of 9 students taking a statistics course has a mean IQ of 120. We could now use this information to estimate the value of their population mean. For instance, we could ask what range of values has a 60% probability of including the population mean. This is an example of what statisticians call a **confidence interval**.

Confidence interval – *The range of values that has a known probability of including the population parameter, usually the mean.*

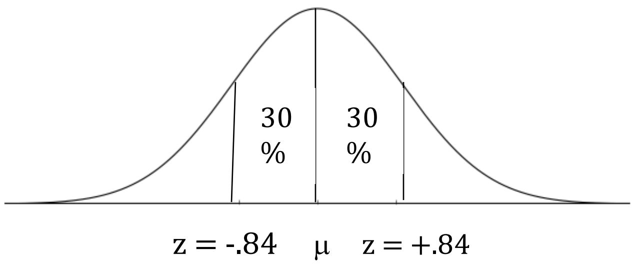
We begin with a figure so we can visualize what we are seeking (Figure 9.7). This is the region of the distribution that is closest to the μ. Thus, for our example 60% of the area of the distribution has been divided into two equal regions, each of 30%, around μ.

Figure 9.7 Illustration of a 60% Confidence Interval



Next, we turn to the z table to ascertain the two values of z that will include 30% of the distribution above and below the mean. The lower value is approximately –0.84. [You can determine this value by looking in the z table (Appendix K, Table 1a) for the z score equivalent to a proportion of 0.20, which is the area in the lower tail. This is found through subtraction: .50, the total proportion of the curve below the mean, minus .30, the proportion already specified.] The upper value is approximately +0.84. (You can determine this value by looking in Table 1b of Appendix K for the z score equivalent of a proportion 0.80, which is 0.50, the area below the mean, plus 0.30, the area indicated above the mean.) Thus we will be looking for an interval that extends from –0.84 to +0.84 standard errors from the population mean. This is shown in Figure 9.8.

Figure 9.8 Identification of the Equivalent z Scores for a 60% Confidence Interval

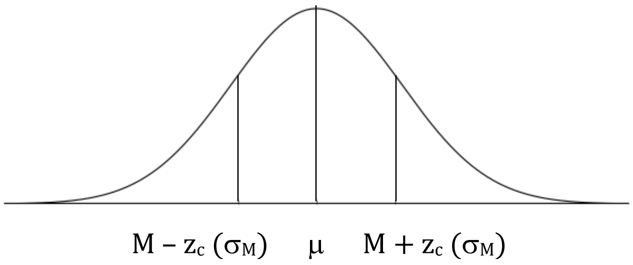


We now need to convert these critical values of z (–0.84 and +0.84) obtained from the z table into the equivalent mean IQ scores based upon samples of size 9. To do so we calculate the following interval:

$$M - z_c\,(\sigma_M) \leq \mu \leq M + z_c\,(\sigma_M)$$

where z_c is the critical value for z we obtained from the z table. Note, however, that when entering values we are ignoring the sign, + or –, of the critical values of z we have just obtained. Thus z_c is an absolute value, in this case 0.84. The lower limit of the confidence interval is $M - z_c\,(\sigma_M)$, and the upper limit is $M + z_c\,(\sigma_M)$. This is shown in Figure 9.9.

Figure 9.9 Determining the Lower and Upper Confidence Interval Limits



In our example the values of M and z_c have already been determined and can be substituted directly into the confidence interval. However, σ_M needs to be calculated from the equation:

$$\sigma_M = \frac{\sigma_X}{\sqrt{n}}$$

The standard deviation of the IQ test is 15, and our sample size is 9, therefore:

$$\sigma_M = \frac{15}{\sqrt{9}}$$

$$= \frac{15}{3}$$

$$= 5$$

Substituting we obtain:

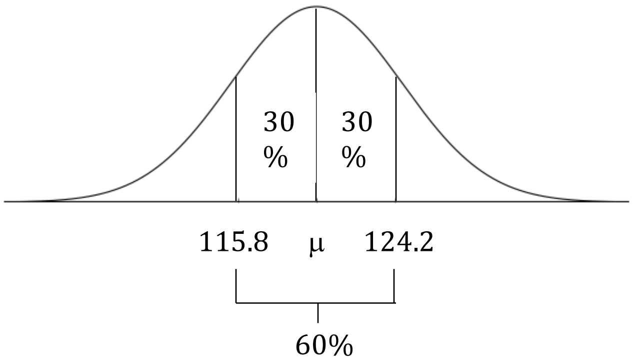
$$120 - 0.84(5) \leq \mu \leq 120 + 0.84(5)$$

$$120 - 4.2 \leq \mu \leq 120 + 4.2$$

$$115.8 \leq \mu \leq 124.2$$

Based upon our sample of size 9 with a mean of 120 and a standard error of 5, we can say that the probability is .60 that a confidence interval with a range of 115.8 to 124.2 will include the population mean IQ of the statistics students *. This interval is illustrated in Figure 9.10.

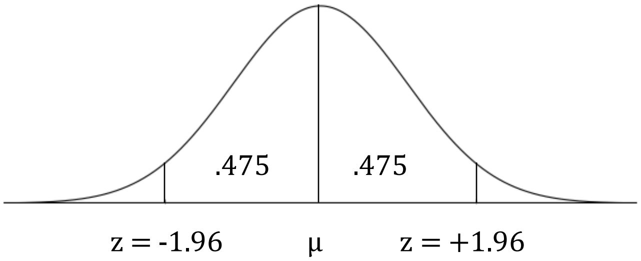
Figure 9.10 Illustration of a 60% Confidence Interval



** The meaning of a confidence interval may be clearer if you look at the situation from two perspectives:*
1. From the standpoint of the population mean. The population mean is either included within the confidence interval, or not. The probability is thus 1 if it is within, and 0 if it does not fall within the confidence interval. And since the population mean is fixed and thus not varying, repeatedly asking if the population mean is within the same confidence interval will lead to the same answer – the probability remains 1 or 0. Thus the probability of the population mean falling within a confidence interval derived from a single sample is either 1 or 0.
2. From the standpoint of the confidence intervals. If we take a series of samples and construct confidence intervals we will find, with alpha equal to .05, that 95% of these confidence intervals will include the population mean, and 5% will not. Thus there is a 95% probability that a particular confidence interval includes the population mean if alpha equals .05. The same logic is used with other values of alpha.

What is the effect if you kept all the other values of our example the same, but changed from a 60% confidence interval to a 95% confidence interval? You would begin, as before, by dividing 0.95 by 2 to obtain 0.475. This is the proportion of the curve desired on each side of the population mean. Recall that the total area on each side of the population mean with a symmetrical distribution is 0.50. The difference between 0.50 and 0.475 is 0.025, which is the area of the curve in each extreme tail. We previously determined that a proportion of 0.025 is equivalent to a z score of –1.96, and a proportion of 0.975 (this is found by adding 0.50, the area below the mean, plus 0.475, the area above the mean) is equivalent to a z score of +1.96. Therefore, we would be looking for an interval that extends from 1.96 SD units below the mean to 1.96 SD units above the mean. This is illustrated in Figure 9.11.

Figure 9.11 Determining a 95% Confidence Interval



The confidence interval would again be calculated as follows:

$$M - z_c\,(\sigma_M) \leq \mu \leq M + z_c\,(\sigma_M)$$

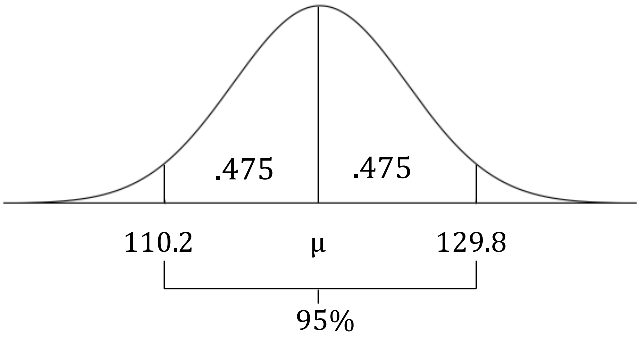
$$120 - 1.96(5) \leq \mu \leq 120 + 1.96(5)$$

$$120 - 9.8 \leq \mu \leq 120 + 9.8$$

$$110.2 \leq \mu \leq 129.8$$

This confidence interval is illustrated in Figure 9.12.

Figure 9.12 Illustration of a 95% Confidence Interval



As you would expect, the 95% confidence interval is considerably larger than the 60% confidence interval we calculated previously (you need a wider range of values to have a .95 probability that a confidence interval will include the population mean than only a .60 probability).

The vast majority of confidence intervals are two-sided, as in the previous examples. It is possible, however, to calculate a one-sided confidence interval when you are making a directional prediction (a one-tailed test). Since this is an introductory text we will limit our discussion of confidence intervals to situations where we have a two-tailed, and thus non-directional, hypothesis.

We have now nearly finished our introduction to the one-sample z test. In closing, we will list the purpose and limitations, and then the assumptions, of the one-sample z test, followed by a brief summary.

Purpose And Limitations Of Using The One-sample z Test

1. *Test for difference.* The one-sample z test is employed to determine whether the difference between a sample mean and an hypothesized or known population mean is due to chance or is instead indicative of a reliable difference. Directional and non-directional hypotheses can be tested.
2. *Does not provide a measure of effect size.* The one-sample z test is a test of significance. It indicates whether or not an outcome is likely to have occurred by chance if the null hypothesis is correct. If the z test is significant a measure of effect size such as Cohen's d should then be calculated. However, as the one-sample z test is not as frequently used as other procedures that will be reviewed in this text a discussion of a measure of effect size for this test was not included.

Assumptions Of The One-sample z Test

1. *Interval or ratio data.* The data are on an interval or ratio scale of measurement.
2. *Random sample.* The sample is drawn at random from the population.
3. *Normally distributed population.* The population from which the sample is drawn has a normal distribution of scores. However, as stated in the Central Limit Theorem, the one-sample z test is accurate (is robust) even if the underlying population is not normally distributed so long as the sample size is at least 30.
4. *The population standard deviation is known.*

Summary Of The One-sample z Test

The one-sample z test is an inferential statistical procedure used to differentiate between null and alternative hypotheses. In order to use the one-sample z test it is assumed that the population from which the sample is drawn is normally distributed and that its standard deviation is known. (The central limit theorem extends the use of the z test to include situations where the shape of the population distribution is not known so long as the sample size is at least 30.) If these conditions are met it is then the case that the sampling distribution of the mean will also be normally distributed and will have a standard error (σ_M) equal to σ_X / $\sqrt{}$n. The one-sample z test can then be conducted and the z table consulted to determine whether to accept the null hypothesis or, instead, to reject the null and accept the alternative hypothesis. Alternatively, a confidence interval can be created.

Progress Check

Assume we are interested in whether an SAT review course actually increases SAT scores. To determine this we randomly select 100 individuals from the general population who then take the

SAT review course. At the conclusion of the course they take the SAT exam. Their mean SAT score is 517. (Recall that the SAT is normally distributed with a mean of 500 and the standard deviation is 100.)

1. Is this a one- or two-tailed test?
2. What is the obtained value for z?
3. What is your decision?

Answers: 1. One-tailed 2. 1.7 3. Reject the null hypothesis

The One-Sample t Test

Although the one-sample z test can be very useful, it is limited to situations in which the value of the population standard deviation is known. As a consequence in many situations it is not possible to utilize a z test. Fortunately, when we do not know the value of the population standard deviation we can estimate it from the sample data. Then we can use a very similar statistical procedure to the one-sample z test that is called the one-sample t test.

In Chapter 3 you learned how to calculate the sample standard deviation. We will see shortly that the equation we have used for the sample standard deviation can serve as the basis for estimating the population standard deviation (σ_X). The symbol for the sample standard deviation used in Chapter 3 was s. However, just as we added the subscript x to the symbol σ to clarify that we are referring to the standard deviation of scores, we will now use s_X in the place of s.

To reacquaint you with the definitions and calculations that you previously learned, the symbols when describing population parameters and sample statistics are presented in Table 9.2 and the equations that were covered in Chapter 3 are shown in Table 9.3.

Table 9.2 Symbols Used when Describing Population Parameters and Sample Statistics

	Population Parameter	Sample Statistic
Size of Data Set	N	n
Mean	μ	M
Variance	σ_X^2	s_X^2
Standard Deviation of Scores	σ_X	s_X

Table 9.3 Equations for the Standard Deviation when Describing Populations and Samples

Population	Sample
$\sigma_X = \sqrt{\frac{\Sigma\,(X - \mu)^2}{N}}$	$s_X = \sqrt{\frac{\Sigma\,(X - M)^2}{n - 1}}$

$\sigma_X = \sqrt{\frac{SS}{N}}$ $\quad\quad$ $s_X = \sqrt{\frac{SS}{n-1}}$

$\sigma_X = \sqrt{\frac{\Sigma x^2}{N}}$ $\quad\quad$ $s_X = \sqrt{\frac{\Sigma x^2}{n-1}}$

It was pointed out in Chapter 3 that the equations for the variance and standard deviation of sample data use n – 1 rather than n in the denominator. Though dividing by n when calculating a standard deviation would provide an accurate measure of sample variability when used as a descriptive statistic, when used as an inferential statistic it consistently underestimates the value of σ_X. In other words, s_X is a **biased estimator** of σ_X if n instead of n – 1 is used in the denominator. A biased estimator does not accurately predict what it is intended to because of systematic error. s_X is biased because a sample consists of a subset of a population and is likely, therefore, not to include the low frequency scores that tend to be more extreme. This becomes an increasingly important issue as the sample size gets smaller. Fortunately, as we have seen, there is a simple solution. Instead of dividing the sum of the squared deviations from the sample mean by the sample size, n, we instead divide by n – 1. When we do this, the systematic error is eliminated. In other words, we actually could compute two different measures of the standard deviation from any sample. One would use n in the denominator and would be appropriate if we were solely interested in the variability of the sample (descriptive statistic). This is a relatively rare situation. The other option uses n – 1 in the denominator and is appropriate if we are interested in using the sample to estimate the variability of the population from which it was chosen (inferential statistic). This is the much more common situation and in order to prevent confusion most texts, including this one, always use n – 1 in the denominator when calculating variances and standard deviations of samples. Thus we do not need to constantly be considering whether we are calculating a standard deviation as a descriptive statistic, or as an inferential statistic. And fortunately the sample mean is an unbiased estimator of the population mean and as a consequence we do not need to consider any correction to its calculation.

Biased estimator – An estimator that does not accurately predict what it is intended to because of systematic error.

By using s_X we have seen that we have an unbiased estimator of σ_X. However, when we use s_X to estimate σ_X we no longer use the one-sample z test. Instead we use the one-sample t test. As the following equations indicate, the one-sample z and t tests are very closely related:

One-sample z Test

$z = \frac{\mathbf{M} - \mu}{\sigma_{\mathbf{M}}}$

where the standard error, $\sigma_M = \frac{\sigma_{\mathbf{X}}}{\sqrt{\mathbf{n}}}$

One-sample t Test

$t = \frac{\mathbf{M} - \mu}{\mathbf{s_M}}$

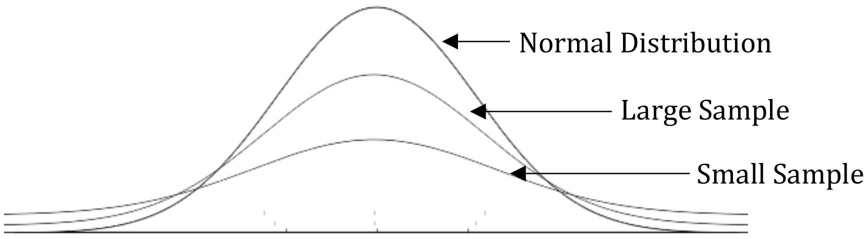
where the standard error, $s_M = \frac{\mathbf{s_X}}{\sqrt{\mathbf{n}}}$

Clearly, the only difference between the equations is that with the t test we use s_M, an *estimate* of the standard error of the population which is derived from the sample data, whereas for the z test we use σ_M, the standard error of the population.

It would be reasonable, but unfortunately incorrect, to assume that we could calculate the t statistic and then enter the z table to find the appropriate proportions. The problem is that while the distribution of z is normal, the shape of the distribution of t varies depending upon the sample size (more precisely upon the degrees of freedom) and is only truly normal when the number of degrees of freedom is infinite. The reason the sample size (degrees of freedom) matters is as follows.

With both the z and t tests we are dividing by a standard error. In the case of the z test, this is σ_M, which is derived from σ_X. In the case of the t test, this is s_M, which is derived from $s_{X.}$ As was noted previously, s_X is an unbiased estimator for σ_X. However, while σ_X is a fixed characteristic of a population, s_X is derived from sample data and thus varies. The consequence is an increased variability in the t distribution, and the smaller the sample the greater the divergence from normal. This, in turn, affects the interpretation of sample data. Specifically, the probabilities found with the normal distribution (used with z) to interpret M – μ differences will be inaccurate if based upon s_M, slightly for moderately sized samples, more dramatically for small samples. And as a result, there is not a single distribution for use with the t test; there is a series or family of distributions, a different distribution for each degree of freedom. When the sample size is small, the difference between the t distribution and the normal distribution (used with z) is substantial. As the sample size increases, the difference between the two distributions becomes smaller. With an infinite sample size the t distribution is normal. Practically speaking, with samples larger than 30 there is little difference between the distributions. The relationship of the family of t distributions to the normal or z distribution is illustrated in Figure 9.13.

Figure 9.13 Relationship Between Sample Size and the t Distribution



From an inspection of Figure 9.13 it is evident that with small sample sizes the t distributions have a greater proportion of their areas located in the extreme tails than occurs with the normal distribution. This means that to include a particular percentage of the curve, such as 95%, we will have to move farther out into the tails. In other words, while we have found from the z table that +/– 1.96 SD from the mean will include 95% of the area of the normal curve, it will be necessary to move farther from the mean to include 95% when we are using a t distribution. How much farther will depend upon the specific sample size or, more precisely, the degrees of freedom. In other words, just as with the chi-square table there will be a different value for each different degrees of freedom. In the case of the t distribution, df = n – 1, where n is the sample size.

Turning to the t table for a two-tailed test (Appendix K, Table 3b), (note the change from z to t tables) we will begin by assuming we have set alpha at .05. Proceeding to the bottom of the column headed by .05 we find a value of 1.96. This is the same value as in the z table and indicates that if the degrees of freedom were infinite the t distribution would be normal and thus the critical value would be the same as with the z distribution. As you go up this column, in other words as the number of degrees of freedom (and thus the sample size) decreases, the critical values of t increase. With 60 degrees of freedom, corresponding to a sample size of 61, the value of t is 2.00. This is only slightly larger than 1.96. However, with 10 degrees of freedom, corresponding to a sample size of 11, the critical value of t has increased to 2.23 and with 1 degree of freedom, corresponding to a sample size of only 2, it has increased dramatically to 12.71. The increase in the size of the critical value for t as the degrees of freedom decreases is the consequence of the shapes of the family of t distributions illustrated in Figure 9.13.

The effect of degrees of freedom on the critical value of the t distribution can be illustrated with an example. Let us assume that we have a sample of 6 subjects randomly selected from a normally distributed population with a μ of 10. The null hypothesis is that the treatment did not have an effect. The alternative hypothesis is that the treatment did have an effect. This is a non-directional (two-tailed) test and we set alpha equal to .05. Following the treatment these subjects have been found to have a M of 14.40 and a SD of 4.90. The standard error, s_M, is therefore:

$$s_M = \frac{\mathbf{s_X}}{\sqrt{\mathbf{n}}}$$

$$= \frac{\mathbf{4.90}}{\sqrt{\mathbf{6}}}$$

$$= \frac{\mathbf{4.90}}{\mathbf{2.45}}$$

$$= 2.00$$

The equation for t is:

$$t = \frac{\mathbf{M} - \mu}{\mathbf{s_M}}$$

Substituting, we have $t = \frac{14.40 - 10}{2.00}$

$$= \frac{\mathbf{4.40}}{\mathbf{2.00}}$$

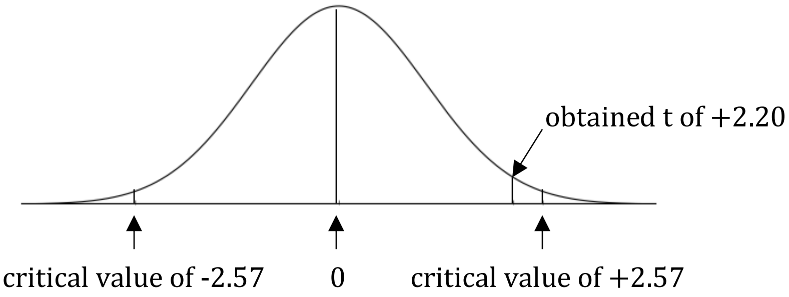
$$= 2.20$$

and

$$\begin{aligned} df &= n - 1 \\ &= 6 - 1 \\ &= 5 \end{aligned}$$

Referring to the t table for a two-tailed test (Appendix K, Table 3b), we find that the critical value with alpha equal to .05 and with 5 degrees of freedom is +/– 2.57. As our obtained t of +2.20 is less than the critical value of +2.57 we retain the null hypothesis. This outcome is illustrated in Figure 9.14. However, it is important to recognize that if the degrees of freedom had been 12 or larger then we would have rejected the null hypothesis.

Figure 9.14 Comparison of Obtained and Critical Values for t



Reporting The Results Of An Insignificant One-Sample t Test

In an article, we would say, "There was insufficient evidence to reject the hypothesis that the sample (M = 14.40, SD = 4.90) was drawn from a population with a mean of 10 (t(5) = 2.20, p > .05)". It is important to note the direction of the > symbol and that no measure of effect size is included since our outcome was not statistically significant.

The Effect Of Increasing Degrees Of Freedom

It was just pointed out that if we had had 12 or more degrees of freedom in the previous example then the outcome would have been statistically significant. Thus, a disadvantage with a small sample size is that you will need a larger experimental effect in order to find a statistically significant difference. In other words, with a small sample size the power of the statistical test is low. This issue is discussed in more detail in Appendix E.

A Measure Of Effect Size For The One-Sample t Test

With the previous example, assuming everything stayed the same except that we had 12 degrees of freedom, then the one-sample t test would be significant which would indicate that the outcome was unlikely to be due to chance. However, as was the case with the one-sample z test, the one-sample t test does not indicate the effect size. Fortunately, the percent of variance explained by the treatment can easily be found by calculating a commonly used measure, **eta squared** (η^2):

$$\eta^2 = \frac{\mathbf{t^2}}{\mathbf{t^2 + df}}$$

For our example assuming there were 12 df this would be:

$$\eta^2 = \frac{\mathbf{2.20^2}}{\mathbf{2.20^2 + 12}}$$

$$= \frac{\mathbf{4.84}}{\mathbf{4.84 + 12}}$$

$$= \frac{\mathbf{4.84}}{\mathbf{16.84}}$$

$$= .287 \text{ or } 28.7\%$$

Thus, in this example with 12 degrees of freedom the treatment would have accounted for 28.7% of the total variance. In an article, after reporting the significant t, we would say, " η^2 equaled .287".

<u>Eta squared</u> (η^2) – A commonly used measure of effect size that indicates the percentage of variation in the dependent variable that is explained or accounted for by the independent variable.

Confidence Interval For t

Just as with z, we can also find a confidence interval for t. And, fortunately, the procedure for finding the confidence interval for a one-sample t statistic is almost identical to the procedure used with z:

$$M - t_c\,(s_M) \leq \mu \leq M + t_c\,(s_M)$$

In this equation, M is the sample mean and t_c is the absolute value of the critical value of t found in the t table (Appendix K, Table 3b). Finally, s_M is the estimate of the population standard error derived from the sample data.

For our example with a sample size of 6, we would have:

$$14.40 - (2.57)\,(2.00) \leq \mu \leq 14.40 + (2.57)\,(2.00)$$

this equals:

$$14.40 - 5.14 \leq \mu \leq 14.40 + 5.14$$

or:

$$9.26 \leq \mu \leq 19.54$$

We would state that with a sample size of 6 (and thus 5 df) there is a 95% probability that a confidence interval with values between 9.26 and 19.54 will include the population mean. (However, recognize that the same clarification that was noted in the discussion of the confidence interval when using z also is true for t.) It is also important to note that in this example the interval includes the hypothetical population mean of 10, which indicates that our t test was not statistically significant. However, if we had used a larger sample the confidence interval would be smaller. If you recalculate the confidence interval, except this time assuming a larger sample size of 13 (and thus 12 df), you will see that the population mean of 10 is no longer included within the confidence interval.

We have now nearly finished our introduction to the one-sample t test. In closing, we will list the purpose and limitations, and then the assumptions, of the one-sample t test, followed by a brief conclusion.

Purpose And Limitations Of Using The One-sample t Test

1. *Test for difference.* With a two-tailed test the null hypothesis is that the treatment does not have an effect. Therefore, any difference between the sample mean and hypothesized population mean is due to chance. The alternative hypothesis is that the treatment does have an effect and, therefore, the difference observed between the two means is not due to chance. The one-sample t test is employed to differentiate between these two hypotheses.
2. *Does not provide a measure of effect size.* The one-sample t test is a test of significance. It indicates whether or not an outcome is likely to have occurred by chance if the null hypothesis is correct. If the t test is significant a measure of effect size such as eta squared (η^2) should then be calculated.

Assumptions Of The One-sample t Test

1. *Interval or ratio data.* The data are on an interval or ratio scale of measurement.
2. *Random sample.* The sample is drawn at random from the population.
3. *Normally distributed population.* The population has a normal distribution of scores. However, as stated in the Central Limit Theorem, the probabilities associated with using the one-sample t test will be accurate even if the underlying population is not normally distributed so long as the sample size is at least 30. If the sample size is less than 30, then the underlying population must be normally distributed. If you cannot collect a larger sample and do not know if the assumption of normality has been met it is best to turn to an alternative test on the same row of Table 9.1.

Conclusion

The one-sample t test is much like the one-sample z test that was covered previously in this chapter. Both assume that the sample is drawn from a normally distributed population. And each can be used to test an hypothesis and to construct a confidence interval. The fundamental difference is that whereas the z test requires that we know the population standard deviation, σ_X, in order to calculate the standard error, σ_M, the t test is more flexible because it uses an estimate of σ_X derived from the sample (s_X). A consequence of using s_X to estimate σ_X is that, particularly with small degrees of freedom, the t distribution differs substantially from the normal distribution. In order to account for this discrepancy there is a family of t distributions.

We have now completed the section of the book dealing with the one-sample z and t tests. Before continuing with the study of additional statistical procedures it may be helpful to take a few moments to review what we have accomplished and to put it into perspective. By referring to Table 9.1 (or Appendix L), you will see that we have begun our review of difference designs for use with interval or ratio data. Specifically, we have learned how to analyze data from a design that uses only one sample. We will soon be discussing more complex designs for use with independent samples as well as designs that use repeated measures. Before doing so, please review Table 9.1 (or Appendix L) in order to once again see the relationships among the statistical procedures.

Glossary Of Terms

Biased estimator – An estimator that does not accurately predict what it is intended to because of systematic error.

Central limit theorem –

–With increasing sample sizes, the shape of the distribution of sample means (sampling distribution of the mean) rapidly approximates the normal distribution irrespective of the shape of the population from which it is drawn.

–The mean of the distribution of sample means (M_G) is an unbiased estimator of the population mean.

–And the standard deviation of the distribution of sample means (σ_M) will equal $\sigma_X / \sqrt{n}$.

Confidence interval – The range of values that has a known probability of including the population parameter, usually the mean.

Error – An outcome due to chance.

Eta squared (η^2) – A commonly used measure of effect size that indicates the percentage of variation in the dependent variable that is explained or accounted for by the independent variable.

Grand mean (M_G) – The mean of the sample means.

Law of large numbers – The larger the sample size, the better the estimate of population *parameters such as μ.*

One-sample t test – An inferential procedure for comparing a sample mean with a population mean when the population standard deviation is not known.

One-sample z test – An inferential procedure for comparing a sample mean with a population mean when the population standard deviation is known.

One-tailed or directional test – An analysis in which the null hypothesis will only be rejected if an extreme outcome occurs in the predicted direction. In such a test, the single area of rejection is equal to alpha and it is located in one tail of the sampling distribution.

Two-tailed or nondirectional test – An analysis in which the null hypothesis will be rejected if an extreme outcome occurs in either direction. In such a test, the area of rejection is divided into two parts, each equal to α / 2.

Sampling distribution of the mean – A theoretical probability distribution of sample means. The samples are all of the same size and are randomly selected from the same population.

Standard error of the mean (SEM) – The standard deviation of the sampling distribution of means.

References

Field, A. (2009). *Discovering statistics using SPSS, 3rd edition.* Los Angeles: SAGE.

Lumley, T., Diehr, P., Emerson, S. & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health, 23*, 151-169.

Questions – Chapter 9

(Answers are provided in Appendix J.)

1. Which of the following is *not* true of the sampling distribution of the mean?
   a. distribution is symmetrical
   b. distribution is normal
   c. distribution is uni–modal
   d. distribution is skewed

2. In a _____ all of the region of rejection is placed in one end of the distribution.
   a. two-tailed test
   b. one-tailed test
   c. non-directional test
   d. none of the above

3. An experimenter wants to test whether a particular intervention will change students' grades. This is an example of a _____ test.
   a. one-tailed
   b. two-tailed
   c. three-tailed

4. In a one-tailed test, the area of rejection is _____.

a. placed in one tail, but it does not matter which tail
b. divided equally in the two tails
c. placed in one tail based upon the previous literature

5. The one-sample t test is used instead of the one-sample z test when _____ is *not* known.
a. The population mean
b. The population size
c. The population standard deviation

6. The outcome of the one-sample z test is a number measured in _____.
a. The units of the original data, such as meters or pounds.
b. Standard error units
c. Units that vary with each problem
d. None of the above

7. With a t test, the _____ must be estimated.
a. Population mean
b. Population size
c. Population standard deviation

8. Degrees of freedom are used with the _____.
a. One-sample z test
b. One-sample t test
c. Both the one-sample z and t tests

9. If we compare the z and t tables, assuming the criterion remains .05 the critical value for z will be _____ the critical value of t.
a. Smaller than
b. Larger than
c. The same as

10. The statement that "the shape of the distribution of sample means (sampling distribution of the mean) rapidly approximates the normal distribution irrespective of the shape of the population from which it is drawn" is a part of the definition of _____.
a. a confidence interval
b. degrees of freedom
c. the central limit theorem
d. the law of large numbers

11. The mean of a large sample provides a better estimate of the population mean than the mean of a small sample. This is an example of _____.
a. a confidence interval
b. degrees of freedom
c. the central limit theorem
d. the law of large numbers

12. Another name for the standard deviation of the sampling distribution of means is the _____.
a. ultimate standard deviation
b. positive standard deviation
c. standard error
d. maximum standard error

13. The Confidence Interval has a known probability of including the _____.

a. population parameter
b. sample statistic
c. either the population parameter or the sample statistic

For questions 14 to 19, use the following information: A researcher is interested in whether drinking orange juice will have an effect on IQ. Accordingly, 25 randomly selected subjects drink orange juice and the group is subsequently found to have a mean IQ of 102. Is this a sufficient difference to conclude that drinking orange juice affects IQ? Set alpha equal to .05. (Remember, the mean of the commonly used IQ tests is 100 and the standard deviation is 15.)

14. What is the null hypothesis?
a. orange juice affects IQ
b. orange juice does not affect IQ

15. What is the standard error?
a. 15
b. 3
c. 9
d. 1

16. What is the value of z for these data?
a. 0.67
b. 1.0
c. 1.8
d. 3.0

17. Is this a one- or two-tailed test?
a. one-tailed test
b. two-tailed test

18. What is the critical value from the z table?
a. 1.96
b. 2.58
c. 1.64 or 1.65
d. 1.00

19. What conclusion do you make?
a. retain the null hypothesis – there is not sufficient evidence to conclude that orange juice affects IQ
b. reject the null hypothesis – there is sufficient evidence to conclude that orange juice affects IQ
c. there is not sufficient evidence to come to any decision

For questions 20 to 25, use the following information: A researcher is interested in whether giving rats a particular diet will have an effect on how fast they run a maze. Accordingly, 20 randomly selected subjects are given the experimental diet and this group is subsequently found to have a mean run time of 130 seconds. Is there sufficient evidence to conclude that the diet has had an effect on running speed if it is known that the mean of the population of rats without the special diet is 125 seconds and the standard deviation of the population is 10.5 seconds? Use alpha equal to .05.

20. What is the null hypothesis?
a. The experimental diet affects running speed

b. The experimental diet does not affect running speed

21. What is the standard error?
    a. 10.5
    b. 20
    c. 4.47
    d. 2.35

22. What is the value of z for these data?
    a. 5
    b. 1.96
    c. 2.13
    d. 3.0

23. Is this a one- or two-tailed test?
    a. one-tailed test
    b. two-tailed test

24. What is the critical value from the z table?
    a. 1.96
    b. 2.58
    c. 1.64 or 1.65
    d. 1.00

25. What conclusion do you make?
    a. retain the null hypothesis – there is not sufficient evidence to conclude that the diet affects running speed
    b. reject the null hypothesis – there is sufficient evidence to conclude that the diet affects running speed
    c. there is not sufficient evidence to come to any decision

For questions 26 to 33, use the following information: A researcher is interested in whether applying a particular fertilizer will have an effect on crop production. Accordingly, 20 randomly selected plants are given the experimental fertilizer and this group is subsequently found to produce a mean of 30 pounds of fruit. Is there sufficient evidence to conclude that the fertilizer has had an effect if it is hypothesized that the mean production without the fertilizer is 25 pounds and the estimate of the population *standard error*, derived from the sample, is 2.13 pounds? Use alpha equal to .05.

26. What is the null hypothesis?
    a. The experimental fertilizer affects crop production
    b. The experimental fertilizer does not affect crop production

27. Would you employ a z or a t test for these data?
    a. z test
    b. t test

28. What is the value of the statistical test for these data?
    a. 2.35
    b. 1.96
    c. 2.19
    d. 3.06

29. Is this a one- or two-tailed test?
   a. one-tailed test
   b. two-tailed test

30. How many degrees of freedom are there?
   a. 18
   b. 19
   c. 20
   d. 21
   e. It is not possible to determine the degrees of freedom for these data.

31. What is the critical value from the appropriate table, assuming alpha is set at .05?
   a. 1.96
   b. 2.58
   c. 2.093
   d. 3.441

32. What conclusion do you make?
   a. retain the null hypothesis – there is not sufficient evidence to conclude that the fertilizer affects crop production
   b. reject the null hypothesis – there is sufficient evidence to conclude that the fertilizer affects crop production
   c. there is not sufficient evidence to come to any decision

33. What is the confidence interval that has a 95% probability of including the population mean?
   a. $24.77 \leq \mu \leq 35.23$
   b. $28.52 \leq \mu \leq 31.48$
   c. $26.54 \leq \mu \leq 33.46$
   d. $20.15 \leq \mu \leq 39.85$
   e. $25.54 \leq \mu \leq 34.46$

For questions 34 - 36 assume that a sample of 16 students taking a statistics course has a mean IQ of 108. What interval has a 95% probability of including the population mean? Remember, $\sigma_X =$ 15 for the IQ test.

34. What is the standard error?
   a. 3.75
   b. 15
   c. 4
   d. 16

35. What is the critical value from the z table?
   a. 1.96
   b. 2.58
   c. 1.64 or 1.65
   d. .67

36. What is the confidence interval that has a 95% probability of including the population mean?
   a. $105.49 \leq \mu \leq 110.51$
   b. $103.5 \leq \mu \leq 112.5$
   c. $100.65 \leq \mu \leq 115.35$

d. $100 \leq \mu \leq 116$

For questions 37 – 39 use the same information as in the previous questions (34 – 36) except answer what interval has a 50% probability of including the population mean.

37. What is the standard error?

a. 3.75
b. 15
c. 4
d. 16

38. What is the critical value from the z table?

a. 1.96
b. 2.58
c. 1.64 or 1.65
d. .67

39. What is the confidence interval that has a 50% probability of including the population mean?

a. $105.49 \leq \mu \leq 110.51$
b. $103.5 \leq \mu \leq 112.5$
c. $100.65 \leq \mu \leq 115.35$
d. $100 \leq \mu \leq 116$

SPSS procedures are rarely use for the statistical tests described in this chapter.

Chapter 10
Finding Differences with Interval and Ratio Data – II: The Independent Samples t and Dependent Samples t Tests

"Science is simply common sense at its best, that is, rigidly accurate in observation, and merciless to fallacy in logic."
Thomas Huxley

Introduction

This chapter will discuss two of the most commonly employed statistical tests. Both are conceptually similar and, as you will see, each is closely related to the one-sample t test that was reviewed in Chapter 9. The first of these tests is called the **independent samples t test**. After discussing this procedure we will turn to the dependent samples t test. These tests are underlined in Table 10.1.

Independent samples t test – An inferential procedure for comparing two means from unrelated samples.

Table 10.1 Overview of Inferential Statistical Procedures for Finding if there is a Difference

	Type of Data			
	Nominal (Frequency)		Ordinal (Ranked)	Interval/Ratio (Continuous Measure)
Research Design		Research Design		
One Variable With At Least Two Outcomes	Goodness-of-fit Chi-Square	One IV With One Sample		One-sample z Test or One-sample t Test
		One IV With Two Or More Independent Samples	*Kruskal–Wallis H*	One-way Between–Subjects ANOVA (Only two independent samples, Independent Samples t Test)
		One IV With One Sample Having Two Or More Repeated Measures		One-way Within–Subjects ANOVA (Only two repeated measures, Dependent Samples t Test)

Two Variables, Each With At Least Two Outcomes	Chi-Square Test of Independence	Two IV Each With Two Or More Independent Samples	Two-way Between–Subjects ANOVA

The italicized procedure is reviewed in Appendix A.

Independent Samples t Test

We often see studies that compare one group of subjects that receives a treatment with another group that serves as a control and does not receive the treatment. For instance, we could study the effect of background noise on learning and memory by randomly assigning subjects to either read a passage in a quiet classroom or while listening to loud music. Subsequently we could compare their retention of the passage's information. Alternatively, two preexisting groups, such as men and women, might be compared on some measure, such as their reaction to bright light. These are both examples of the two independent samples design. Studies that compare two independent groups are very popular and, if the data are interval or ratio, and assumptions are met, can be analyzed with a form of the t test that is called the independent samples t test, the t test for independent samples, or just the independent t test.

Table 10.1 indicates that with this experimental design and ordinal data we would use the Kruskal-Wallis H test which is reviewed in Appendix A (there are other alternatives). With interval or ratio data either the independent samples t test or the one-way between-subjects ANOVA (analysis of variance, reviewed in Chapter 11) is appropriate. The advantage of the independent samples t test compared to the ANOVA is that it is somewhat easier to calculate. However, the one-way between-subjects ANOVA is more flexible as it can be used with designs that have more than two samples.

As was just noted, the t test for two independent samples (independent samples t test) is conceptually very similar to the t test for one sample of subjects which was reviewed in Chapter 9. As you recall, when there is only one sample, the t test examines whether a difference between the sample mean (M) and the population mean (μ) is likely to have happened by chance. This is accomplished by converting this difference into standard deviation units. And what distinguishes the one-sample t test from the one-sample z test is that with the t test we do not need to know the value of the population standard deviation. Instead we substitute an estimate of the population standard deviation (σ_X) derived from the sample (s_X). This estimated standard deviation is then used to calculate the standard error of the mean (s_M). Dividing the difference between the sample mean and the population mean ($M - \mu$) by s_M leads to an outcome measured in standard deviation units. Specifically, for a one-sample study:

$$t = \frac{\mathbf{M} - \mu}{\mathbf{s_M}}$$

where the standard error, $s_M = \frac{\mathbf{s_X}}{\sqrt{\mathbf{n}}}$

As there are a series of t distributions, the degrees of freedom, which for the one-sample t test are equal to n – 1, must then be calculated. Finally, the value obtained for t is compared with the critical value found in the t table (Appendix K, Tables 3a and b).

With the independent samples t test the logic is essentially the same. However, as we are now dealing with two sample means, not one, we are no longer examining whether a difference between a single sample mean and a population mean (M – μ) is likely to have happened by chance. Instead we are comparing the difference of a pair of sample means ($M_1 - M_2$) to the hypothesized difference between the corresponding pair of population means ($\mu_1 - \mu_2$). More specifically, we are examining whether this difference [$(M_1 - M_2) - (\mu_1 - \mu_2)$] is likely to have happened by chance. To do so we substitute a *difference between means* for each *mean* in the above one-sample t equation. Thus instead of a single sample mean (M) we substitute the difference between sample means ($M_1 - M_2$). And instead of a single population mean (μ) we substitute the difference between two population means ($\mu_1 - \mu_2$). Finally, instead of dividing by a standard error *of the mean* (s_M), we substitute the **standard error of the difference between sample means** ($\mathbf{s}_{(\mathbf{M_1} - \mathbf{M_2})}$). (Note that $\mathbf{s}_{(\mathbf{M_1} - \mathbf{M_2})}$ is a single number.) Thus, for an independent samples t test the equation becomes:

$$t = \frac{(\mathbf{M_1} - \mathbf{M_2}) - (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})}{\mathbf{s}_{(\mathbf{M_1} - \mathbf{M_2})}}$$

Standard error of the difference between sample means ($\mathbf{s}_{(\mathbf{M_1} - \mathbf{M_2})}$) – The standard deviation of the sampling distribution of the difference between sample means.

The parallels between the one-sample t test and the independent samples t test may become clearer by referring to Table 10.2.

Table 10.2 Parallels Between the One Sample t Test and the Independent Samples t Test

One Sample t Test		Independent Samples t Test	
Sample mean	M	Difference between sample means	$M_1 - M_2$
Population mean	μ	Hypothesized difference between population means	$\mu_1 - \mu_2$
Standard error of the mean	s_M	Standard error of the difference between sample means	$\mathbf{s}_{(\mathbf{M_1} - \mathbf{M_2})}$
Equation for t	$t = \frac{\mathbf{M} - \mu}{\mathbf{s_M}}$		$t = \frac{(\mathbf{M_1} - \mathbf{M_2}) - (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})}{\mathbf{s}_{(\mathbf{M_1} - \mathbf{M_2})}}$

As you will see, the equation for the independent samples t test is easy to use. However, before turning to an example we will first review what this equation is accomplishing and then turn to an explanation of the logic underlying this form of the t test.

What The Equation For The Independent Samples t Test Accomplishes

The numerator of the independent samples t test indicates that we are to find the difference between our two sample means and from this subtract the hypothesized difference between the corresponding population means. Finding sample means is not difficult and, clearly, neither is finding their difference.

The proposed difference between the two population means is a reflection of the null hypothesis. For instance, if the null hypothesis is that the treatment will not have an effect, then the control and experimental population means are assumed to be equal and thus their difference ($\mu_1 - \mu_2$) is predicted to be zero. In this case, which is quite common, the numerator of the equation for t reduces to simply the difference between the two sample means. On the other hand, the null hypothesis might state that $\mu_1 - \mu_2$ is not zero. For instance, previous research may suggest that there is a pre-existing difference between two groups. An example would be that men are generally a few inches taller than women, and thus the hypothesized difference between the heights of men and women would not be zero.

The difference between the two sample means and the two population means is then divided by the appropriate standard deviation measure. In the case of the independent samples t test this is the standard error of difference between sample means ($\mathbf{s}_{(\mathbf{M_1} - \mathbf{M_2})}$). This will convert the difference found in the numerator into standard deviation units. Then the outcome is interpreted by referring to the t table.

The Logic Of The Independent Samples t Test

Before turning to an example it is important to understand how the independent samples t test is related to previous procedures we have discussed. We will begin with a review of the logic of the z score and then turn to the one-sample z and one-sample t tests as these form the basis for the independent samples t test. This discussion is somewhat theoretical, but it is useful in understanding the logic of these statistical tests.

In Chapter 4 it was noted that to convert a score (X) into a z score we use the equation $z = (X - \mu) / \sigma_X$. In this equation we are dividing the difference $(X - \mu)$ by a standard deviation (σ_X) in order to convert this difference into standard deviation units. If the distribution of the population of scores (X) is normal we can then use the z table to determine the probabilities associated with this difference.

The logic for the one-sample z *test* is the same: we take a difference, this time $(M - \mu)$, and divide by a standard deviation. However, as we are now dealing with a sample mean (M) rather than a score (X), we need to divide by a measure of the variability of sample means. This measure of variability is called the standard error of the mean and has the symbol σ_M. Thus the equation for a *z test* becomes $z = (M - \mu) / \sigma_M$. Since it is known that the theoretical distribution of sample means, which is called the sampling distribution of the mean, is normal when the population of scores from which the sample was drawn was normal, we can then turn to the z table and compare our outcome to the critical value.

Recall that to conduct a z test we need to know the standard deviation of the population (σ_X) from which the sample was drawn. We can then calculate the standard error of the mean, as $\sigma_M = \sigma_X / \sqrt{n}$. Unfortunately, in most cases we do not know σ_X and thus we cannot find the value of σ_M. However, in Chapter 9 you learned that σ_M can be estimated from the variability in the sample (s_X). Specifically, $s_M = s_X / \sqrt{n}$. The equation for the one-sample t test is $t = (M - \mu) / s_M$ which is identical to the equation for the one-sample z test except that we are using s_M as an estimate of σ_M. Finally, we do not use the z table with a t test but instead turn to the t table and take the sample size into account by calculating the degrees of freedom.

The logic for the independent samples t test (two independent samples t test) parallels what was just said for the one-sample t test. Once again we have a difference being converted into standard deviation units. And once again we must use the appropriate estimate of the variability to do so. Finally, we must utilize the t table rather than the z table and take into account the degrees of freedom. What is new is that we are now comparing the difference of two sample means $(M_1 - M_2)$ to the predicted difference of the corresponding population means $(\mu_1 - \mu_2)$. And to convert this difference of sample and population means into standard deviation units we must divide by an appropriate measure of variability, in this case the standard error of the difference between sample means $(\mathbf{s}_{(\mathbf{M_1} - \mathbf{M_2})})$. The result, as was noted previously, is that the equation for the independent samples t test closely parallels the equation for the one-sample t test.

And it was noted previously in the discussion of the logic of the z test that if the population of scores from which the sample was drawn was normally distributed, then the sampling distribution of the mean would also be normally distributed. Similarly, it can be shown that the sampling distribution of the difference between sample means is also normally distributed if each of the populations from which the samples are drawn are normally distributed. However, because we are utilizing an estimate of the standard deviation of the difference between means which is derived from the samples, we must once again use the t table rather than the z table. Thus, the logic of the independent samples t test directly parallels the logic of the one-sample t test.

An example illustrating the calculation of an independent samples t test will assist you in seeing the parallel between this statistical procedure and the one-sample t test which was discussed in Chapter 9.

Conducting An Independent Samples t Test

An example of a two independent samples design would be a fictitious comparison of the effectiveness of two methods for teaching statistics. To conduct the study, each subject would be randomly assigned to either the standard teaching procedure or an alternative procedure (the sample sizes do not have to be equal but they should be similar – and in this example very small samples were chosen to aid in the calculations). The standard procedure would be considered the control condition, and the alternative procedure would be the experimental condition. As is often the case, the null hypothesis is that there is no difference between the conditions. The alternative hypothesis is that there is a difference. After exposure to either the standard or alternative teaching procedure each subject would then be tested to determine their mastery of statistics. This is a two-tailed test and assuming that alpha was set to .05, what should the researcher decide about the effectiveness of the teaching procedures if the scores in Table 10.3 were obtained on a 15-item quiz?

Table 10.3 Example 1: Two Samples of Ratio Data and the Initial Calculations

Experimental Condition			Control Condition		
X_1	$(X_1 – M_1)$	$(X_1 – M_1)^2$	X_2	$(X_2 – M_2)$	$(X_2 – M_2)^2$
13	3	9	6	2	4
12	2	4	6	2	4
11	1	1	4	0	0
9	–1	1	2	–2	4
8	–2	4	$\underline{2}$	$\underline{–2}$	$\underline{4}$
$\underline{7}$	$\underline{–3}$	$\underline{9}$			
$\Sigma X_1 = 60$	$\Sigma x_1 = 0$	$\Sigma x_1^2 = 28$	$\Sigma X_2 = 20$	$\Sigma x_2 = 0$	$\Sigma x_2^2 = 16$
$n_1 = 6$			$n_2 = 5$		
$M_1 = 60 / 6$			$M_2 = 20 / 5$		
$= 10$			$= 4$		

Recall that the equation for t with two independent samples was given previously as:

$$t = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{s_{(M_1 - M_2)}}$$

As the difference between the population means ($\mu_1 - \mu_2$) in our example was hypothesized to be zero, the equation for t becomes:

$$t = \frac{(M_1 - M_2) - 0}{s_{(M_1 - M_2)}}$$

Substituting we have:

$$= \frac{(10 - 4) - 0}{s_{(M_1 - M_2)}}$$

$$= \frac{6}{s_{(M_1 - M_2)}}$$

To find the value of the denominator of this equation, $\mathbf{s}_{(M_1 - M_2)}$ (the standard error of the difference between sample means), we first must find the values of the variance for each sample ($\mathbf{s}^2_{X_1}$ and $\mathbf{s}^2_{X_2}$). The variance of the experimental group (Condition 1) is:

$$s^2_{X_1} = \frac{\Sigma(X_1 - M_1)^2}{n_1 - 1}$$

$$= \frac{\Sigma x_1^2}{n_1 - 1}$$

$$= \frac{28}{6 - 1}$$

$$= \frac{28}{5}$$

$= 5.60$ (Thus the SD = 2.37)

For the control group (Condition 2) the variance is:

$$s^2_{X_2} = \frac{\Sigma(X_2 - M_2)^2}{n_2 - 1}$$

$$= \frac{\Sigma x_2^2}{n_2 - 1}$$

$$= \frac{16}{5 - 1}$$

$$= \frac{16}{4}$$

$= 4.00$ (Thus the SD = 2.00)

The standard error of the difference between sample means is found with the following particularly impressive looking equation. (This equation works for samples with equal or unequal sample sizes.) Fortunately, as you will see, it is easy to use:

$$s_{(M_1 - M_2)} = \sqrt{}[\frac{(n_1 - 1)\, s^2_{X_1} + (n_2 - 1)\, s^2_{X_2}}{n1 + n2 - 2}(\frac{1}{n_1} + \frac{1}{n_2})]$$

$$= \sqrt{}[\frac{(6 - 1)(5.60) + (5 - 1)(4.00)}{6 + 5 - 2}(\frac{1}{6} + \frac{1}{5})]$$

$$= \sqrt{}[\frac{(5)(5.60) + (4)(4.00)}{11 - 2}(\mathbf{0.17 + 0.20})]$$

$$= \sqrt{}[\frac{28 + 16}{9}(\mathbf{0.37})]$$

$= \sqrt{[(4.89)(0.37)]}$

$= \sqrt{1.81}$

$= 1.35$

The value for t is therefore:

$$t = \frac{\mathbf{6}}{\mathbf{s}_{(\mathbf{M_1} - \mathbf{M_2})}}$$

$$= \frac{\mathbf{6}}{\mathbf{1.35}}$$

$$= 4.44$$

where $df = n_1 + n_2 - 2$

$= 6 + 5 - 2$

$= 11 - 2$

$= 9$

The critical value of t obtained from the t table for a two-tailed test with 9 df and alpha equal to .05 is found to be 2.26 (Appendix K, Table 3b). Thus we will reject the null hypothesis if our calculated t is either less than –2.26 or greater than +2.26. As our calculated value of +4.44 is more standard deviation units from the mean than is the critical value of +2.26 we reject the null hypothesis that the control and experimental groups come from populations with equal means and accept the alternative hypothesis that the population means are different. In other words, we conclude that the treatment had an effect. More specifically, we note that as the mean score for the experimental condition is greater than the mean score for the control condition there is evidence that the alternate teaching procedure increased students' scores.

Of course, we still do not know how large the effect was. In order to ascertain the percent of variance explained by the treatment we once again calculate eta squared (η^2) (there are other options). For the independent samples t test, eta squared is found with the same equation as is used with the one-sample t test:

$$\eta^2 = \frac{\mathbf{t^2}}{\mathbf{t^2 + df}}$$

For our example with 9 df, this would be:

$$\eta^2 = \frac{\mathbf{4.44^2}}{\mathbf{4.44^2 + 9}}$$

$$= \frac{\mathbf{19.71}}{\mathbf{19.71 + 9}}$$

$$= \frac{\mathbf{19.71}}{\mathbf{28.71}}$$

$$= .69 \text{ or } 69\%$$

Thus, in this example the treatment accounted for 69% of the total variance, which is a very large effect.

Confidence Interval For An Experiment That Could Have Utilized The Independent Samples t Test

With an independent samples design a researcher could determine a confidence interval either instead of conducting the t test or as a supplement to the t test.

The procedure for finding the confidence interval for a two-sample t (independent samples t) statistic is similar to what was used with a one-sample t:

For one-sample t:

$$M - t_c\,(s_M) \leq \mu \leq M + t_c\,(s_M)$$

For two-sample t (independent samples t):

$$[(M_1 - M_2) - t_c\,(\mathbf{s}_{(\mathbf{M_1} - \mathbf{M_2})})] \leq (\mu_1 - \mu_2) \leq [(M_1 - M_2) + t_c\,(\mathbf{s}_{(\mathbf{M_1} - \mathbf{M_2})})]$$

For our example we would have:

$$[(10 - 4) - (2.26)(1.35)] \leq (\mu_1 - \mu_2) \leq [(10 - 4) + (2.26)(1.35)]$$

This equals:

$$6 - 3.05 \leq (\mu_1 - \mu_2) \leq 6 + 3.05$$

$$2.95 \leq (\mu_1 - \mu_2) \leq 9.05$$

We would state that with 9 df there is a 95% probability that a confidence interval with values between 2.95 and 9.05 will include the difference between the experimental population mean and the control population mean. (However, refer to the clarification included with the discussion of the confidence interval when using z.)

Reporting The Results Of An Independent Samples t Test

In an article until recently we would have reported, "There was sufficient evidence to reject the null hypothesis that the teaching techniques were equivalent. Mastery of statistics was found to be greater in the experimental condition (*M* = 10, *SD* = 2.37) than in the control condition (*M* = 4, *SD* = 2.00) ($t(9) = 4.44$, $p < .05$, $\eta^2 = .69$)". At the end of this chapter we redo this problem using SPSS. This allows us to give a more accurate value for t, and provides the p-value and confidence interval. Using SPSS we would now report, ($t(9) = 4.48$, $p = .002$, $\eta^2 = .69$, 95% *CI* [2.97, 9.03]. Note that the p-value of .002 is less than our α of .05, confirming that we would reject the null hypothesis and that our calculations by hand were accurate except for minor rounding error.

Summary To This Point

To summarize to this point, the two-sample t (independent samples t) test is much like the one-sample t test. Both assume that the sample(s) is (are) drawn from a normally distributed population(s). And each t test is flexible because, unlike the z test, each uses an estimate derived from the sample(s) to determine the necessary standard error. Finally, the one-sample z test, one-sample t test and the two-sample t test can be used to test an hypothesis and/or to construct a confidence interval.

Second Example Of An Independent Samples t Test

In the previous example of the t test each subject was randomly assigned to either the control or the experimental group. The t test is also commonly used when the subjects cannot be randomly assigned. For instance, a researcher might want to study why there are fewer women than men engineers. One way to examine this question would be to determine how attractive engineering fields are to men and women. Let us assume that an initial study found that men rated engineering fields as being 10 points more attractive on some measure than women did. The researcher might then want to determine the effect of an intervention designed to increase women's interest in engineering. In this hypothetical study there would be two groups, men and women. Obviously, however, a subject cannot be randomly assigned to be either a man or a woman. A subject comes to the experiment already being a man or a woman. Nevertheless, a t test can be used to analyze the results.

Specifically, in this example let us assume that the researcher wanted to test the effectiveness of an intervention for women consisting of a talk, several readings and a meeting with a successful woman engineer. The null hypothesis is that the intervention would not decrease the difference between men and women and, therefore, the women would continue to rate engineering fields as being 10 points less attractive than men do. The alternative hypothesis is that the intervention would decrease the difference in the ratings of interest in engineering. This is a one-tailed hypothesis as a directional prediction is being made. As usual, we assume that alpha was set to .05. The researcher planned to include equal numbers of men and women but, as is often the case, several subjects dropped out of the study for various reasons. As a result at the end of the study there were only 7 women and 5 men. (Note that these very small samples were chosen to simplify the calculations.) Their hypothetical ratings of the attractiveness of engineering, along with the initial calculations, are listed in Table 10.4.

Table 10.4 Example 2: Using the t Test with Nonrandom Assignment of Subjects

	Men			Women	
X_1	$(X_1 - M_1)$	$(X_1 - M_1)^2$	X_2	$(X_2 - M_2)$	$(X_2 - M_2)^2$
90	13.4	179.56	87	14.71	216.38

82	5.4	29.16	80	7.71	59.44
76	–0.6	0.36	76	3.71	13.76
72	–4.6	21.16	74	1.71	2.92
63	–13.6	184.96	70	–2.29	5.24
			65	–7.29	53.14
			54	–18.29	334.52
$\Sigma X_1 = 383$	$\Sigma x_1 = 0$	$\Sigma x_1^2 = 415.20$	$\Sigma X_2 = 506$	$\Sigma x_2 = 0$	$\Sigma x_2^2 = 685.40$
$n_1 = 5$			$n_2 = 7$		
$M_1 = 383 / 5$			$M_2 = 506 / 7$		
$= 76.60$			$= 72.29$		

Recall that the equation for the independent samples t test is:

$$t = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{s_{(M_1 - M_2)}}$$

where $s_{(M_1 - M_2)}$ is the standard error of the difference between sample means.

As the difference between the population means was hypothesized to be 10 points this equation becomes:

$$t = \frac{(M_1 - M_2) - 10}{s_{(M_1 - M_2)}}$$

Substituting we have:

$$= \frac{(76.60 - 72.29) - 10}{s_{(M_1 - M_2)}}$$

$$= \frac{4.31 - 10}{s_{(M_1 - M_2)}}$$

$$= \frac{-5.69}{s_{(M_1 - M_2)}}$$

To find the value of the standard error, $s_{(M_1 - M_2)}$, we must first find the variances $s^2_{X_1}$ and $s^2_{X_2}$. The estimate of the population variance for the men (Group 1) is:

$$s^2_{X_1} = \frac{\Sigma x_1^2}{n_1 - 1}$$

$$= \frac{415.20}{5 - 1}$$

$$= \frac{415.20}{4}$$

$$= 103.80$$

For the women (Group 2) the estimate of the population variance is:

$$s^2_{X_2} = \frac{\Sigma x_2^2}{n_2 - 1}$$

$$= \frac{685.40}{7 - 1}$$

$$= \frac{685.40}{6}$$

$$= 114.23$$

We can now find the standard error of the difference between sample means:

$$s_{(M_1 - M_2)} = \sqrt{[}\ \frac{(n_1 - 1)\, s^2_{X_1} + (n_2 - 1)\, s^2_{X_2}}{n1 + n2 - 2}\ (\frac{1}{n_1} + \frac{1}{n_2})]$$

$$= \sqrt{[}\frac{(5-1)(103.80) + (7-1)(114.23)}{5+7-2}\ (\frac{1}{5} + \frac{1}{7})]$$

$$= \sqrt{[}\frac{(4)(103.80) + (6)(114.23)}{10}(0.20 + 0.14)]$$

$$= \sqrt{[}\frac{415.20 + 685.38}{10}(0.34)]$$

$$= \sqrt{[}(110.06)(0.34)]$$

$$= \sqrt{37.42}$$

$$= 6.12$$

The value for t is therefore:

$$t = \frac{-5.69}{6.12}$$

$$= -0.93$$

where $df = n_1 + n_2 - 2$

$$= 5 + 7 - 2$$

$$= 12 - 2$$

$$= 10$$

From the t table the critical value of a one-tailed t with 10 df and with alpha equal to .05 is found to be 1.81 (Appendix K, Table 3a). As we are predicting a decrease in the difference of the ratings our critical value becomes –1.81. We note that our outcome is in the predicted direction (be careful of the meaning of the obtained t). However, since our obtained value of –0.93 is fewer standard deviation units from the mean than is the critical value, we do not reject the null hypothesis. Instead, we conclude that there is not sufficient evidence that the intervention affected the women's interest in engineering relative to the men's interest. Of course, the researcher should recognize that the samples are much too small. A measure of effect size, such as eta squared, is usually not calculated because a significant outcome was not obtained. Finally, a one-sided confidence interval could be calculated, but this is not common.

Purpose And Limitations Of Using The Independent Samples t Test

1. *Test for difference.* The null hypothesis is usually that the treatment does not have an effect. Therefore, if the null is retained any difference between the sample means is assumed to be due to chance. The alternative hypothesis is that the treatment does have an effect and, therefore, that the two samples are drawn from populations with

different means. The independent samples t test is employed to differentiate between these two hypotheses.

2. *Does not provide a measure of effect size.* The independent samples t test is a test of significance. It indicates whether or not an outcome is likely to have occurred by chance if the null hypothesis is correct. If the t test is significant, a measure of effect size, such as eta squared (η^2), should then be calculated.
3. *Compares two sample means.* The independent samples t test is limited to comparing two sample means.

Assumptions Of The Independent Samples t Test

1. *Interval or ratio data.* The data are on an interval or a ratio scale of measurement.
2. *Random samples.* Each sample is drawn at random from a population.
3. *Data within each treatment level are independent.* The datum from one subject is not affecting the datum from another.
4. *Normally distributed populations.* Each population from which a sample is drawn has a normal distribution of scores. However, as stated in the Central Limit Theorem, the t test will be accurate (is robust) so long as each sample size is at least 30. If a sample size is less than 30 then it is important that the underlying population be normally distributed. If you cannot collect a larger sample and do not know if the assumption of normality has been met, it may be best to convert the data to an ordinal scale and turn to an alternative test on the same row of Table 10.1.
4. *Population variances are equal.* The two populations from which samples are drawn have equal variances. With SPSS, this assumption is examined with **Levene's test of equality of variances** (also called Levene's test of homogeneity of variances).

   <u>Levene's test of equality of variances</u> – Procedure used with SPSS to test the assumption that samples are drawn from populations which have equal variances.

Effect Of Violating The Assumptions

The independent samples t test has been found to be robust. This means that it leads to accurate decisions even when some assumptions are violated. However, if the sample sizes are dramatically unequal, the sample distributions have obviously different shapes or the sample variances are clearly not equal, you should not use the t test. Instead, you might consider converting your interval or ratio data into ordinal data and then turning to an appropriate test for the same experimental design in Table 10.1.

Conclusion

The independent samples t test is a commonly used statistical procedure. It compares two sample means and is easy to calculate. As you will see in Chapter 11, the independent samples t test is a special case of the one-way between-subjects ANOVA and any time you could use the two-tailed independent samples t test you could have used the more flexible one-way between-subjects ANOVA. However, the calculations for the one-way between-subjects ANOVA are somewhat more involved.

Progress Check

In a hypothetical study a researcher is interested in whether taking a motorcycle driving safety course decreases the subsequent number of accidents experienced by motorcycle drivers. To determine this, the researcher checks the driving statistics for a 10-year period for motorcycle drivers who either did, or did not, attend a safety course. The mean number of accidents reported for the drivers who took the course was 1.24. The mean number of accidents reported for the drivers who did not take the course was 1.62. And the standard error of the difference between sample means was 0.30. There were 22 degrees of freedom and the alpha was set at .05.

1. Is this a one- or two-tailed test?
2. What is the value of t?
3. What is your decision?

Answers: 1. One-tailed 2. +/–1.27 (The sign depends upon the order the sample means are entered into the equation.) 3. Accept the null hypothesis.

Dependent Samples t Test

Introduction

You have just learned that the independent samples t test is appropriate for experimental designs that have two independent samples, and one dependent variable which is measured at the interval or ratio level. It was also shown that the independent samples t test was closely related to the one-sample t test that was covered in Chapter 9.

The fundamental difference when using the **dependent samples t test** instead of the independent samples t test is that we no longer have independent samples. Instead, the subjects assigned to each value of the treatment are related or paired in some manner. Most commonly there are **repeated measures** on the same subjects. The dependent samples t test is underlined in Table 10.1.

<u>Dependent samples t test</u> – An inferential procedure for comparing two sample means based upon repeated measures of the same subjects, or measures from pairs of subjects who are related in some way.

<u>Repeated measures design</u> – A research design in which each subject is tested more than once.

The dependent samples t test is also closely related to the one-sample t test. As you recall, when there is only one sample the t test converts a deviation between the sample mean (M) and the population mean (μ) into standard deviation units by dividing the difference between these means by the estimate of the standard deviation of sample means. This estimated standard deviation is called the standard error of the mean (s_M). Specifically, for a study utilizing a one-sample t test:

$$t = \frac{\mathbf{M - \mu}}{\mathbf{s_M}}$$

where the standard error, $s_M = \frac{\mathbf{s_X}}{\sqrt{\mathbf{n}}}$

This standard error could be rewritten as $s_M = \frac{\mathbf{s_X}}{\sqrt{\mathbf{n_X}}}$ to emphasize that n is referring to the number of scores.

The essential difference when employing the dependent samples t test is that instead of considering a mean of a set of scores (M) we are now dealing with the mean of a set of difference scores (M_D). Each of these **difference scores (D)** is based upon either two measurements from the same individual (repeated measures design) or, less commonly, two measurements from pairs of related or matched subjects (**matched subjects design**). It is important to note that in both of these situations the two measurements are of the same dependent variable. The mean of the differences between the measurements (M_D) is then compared to the expected value of this mean (μ_D). The result is that the numerator of the dependent t test ($M_D - \mu_D$) looks quite similar to the numerator of the one-sample t test ($M - \mu$) .

In the one-sample t test the difference between sample and population means obtained in the numerator is then divided by an estimate of the variability of sample means (s_M). This converts the deviation in the numerator into standard deviation units. The outcome is then compared to the critical value obtained from the t table. Similarly, with the dependent samples t test the value obtained in the numerator is then divided by an estimate of the variability of means, but in this case it is the estimate of the variability of means of difference scores, not means of scores as was the case with the one-sample t test. This new measure of variability is called the **standard error of the mean difference** ($\mathbf{s_{M_D}}$). And as we are employing an estimate of this measure of variability derived from the data we must continue to use the t table rather than the z table. Thus, we must also account for the degrees of freedom.

Difference score (D) – The difference between two measurements from the same individual (repeated measures design) or two measurements from pairs of matched subjects (matched subjects design).

Matched subjects design – A research design in which equivalent subjects are paired and then one of the subjects is randomly assigned to each group.

Standard error of the mean difference (s_{M_D}) – The standard deviation of the means of difference scores. More precisely, the standard deviation of the sampling distribution of the means of difference scores.

Of course, use of the t table with the dependent t test assumes that the theoretical frequency distribution of the means of difference scores is normally distributed, just as with the one-sample t test it was assumed that the theoretical frequency distribution of sample means was normally distributed. Fortunately, it is known that so long as the distributions of the two samples of original scores are normal, then the distribution of the means of their differences will also be normal. The logic for the dependent samples t test therefore closely parallels the logic of the one-sample t test. And, not surprisingly, the equation for the dependent samples t test will also look very much like the equation for the one-sample t test that was reviewed in Chapter 9.

The parallels between the one-sample t test and the dependent samples t test may become clearer by referring to Table 10.5.

Table 10.5 Parallels Between the One-Sample t Test and the Dependent Samples t Test

	One Sample t	Dependent Samples t
Sample mean	M (could be written as M_X)	M_D
Population mean	μ (could be written as μ_X)	μ_D
Standard error	s_M (could be written as s_{M_X})	s_{M_D}
Equations for t	$t = \frac{M - \mu}{s_M}$	
	Could be written as $t = \frac{M_X - \mu_X}{s_{M_X}}$	$t = \frac{M_D - \mu_D}{s_{M_D}}$
Relationship of standard error and standard deviation		
	$s_M = \frac{s_X}{\sqrt{n}}$	
	Could be written as $s_{M_X} = \frac{s_X}{\sqrt{n_X}}$	$s_{M_D} = \frac{s_D}{\sqrt{n_D}}$
where		
	s_X = standard deviation of	s_D = standard deviation of

a set of scores	a set of difference scores
$\mathbf{n_X}$ = the number of scores	$\mathbf{n_D}$ = the number of difference scores, which is equal to the number of pairs of scores
df = $\mathbf{n_X}$ – 1	df = $\mathbf{n_D}$ – 1

To summarize to this point, in both the one-sample t test and the dependent samples t test, a measure of difference found in the numerator is being converted into standard deviation units by dividing by a standard error. Then the outcome is interpreted by referring to the t table. Thus, the logic of the dependent samples t test directly parallels the logic of the one-sample t test.

Conducting The Dependent Samples t Test

For an example of a repeated measures study let us assume that you are interested in testing whether a fuel additive will change a car's gas mileage as claimed in an advertisement. One option would be for you to randomly assign each vehicle to either the control (no additive) or experimental (additive) condition, and subsequently use an independent samples t test to compare their mileages. Alternatively, you could compare the mileage of the same vehicles with and without the additive. In this case there would be two measures of fuel economy for each vehicle. If the null hypothesis was that the additive would have no effect, then the population mean mileage without the additive (μ_{WO}) would be predicted to equal the population mean mileage with the additive (μ_W). Thus, the null hypothesis would state that $\mu_W - \mu_{WO} = 0$. The alternative hypothesis would be that $\mu_W - \mu_{WO} \neq 0$. This is a two-tailed test and, as usual, we set $\alpha = .05$. The very small, hypothetical data set and initial computations for our dependent samples t test are shown in Table 10.6.

Table 10.6 Example 1: Repeated Measures Data and Initial Calculations

Vehicle	Mileage	Mileage	Difference		
	With	Without	Scores		
	Additive	Additive	D	$(D - M_D)$	$(D - M_D)^2$
1	13	12	1	0.50	0.25
2	15	13	2	1.50	2.25
3	14	15	–1	–1.50	2.25
4	17	17	0	–0.50	0.25
5	24	20	4	3.50	12.25
6	22	25	–3	–3.50	12.25

$\sum D = 3 \qquad \sum(D - M_D) = 0 \qquad \sum(D - M_D)^2 = 29.50$

$M_D = \frac{\sum \mathbf{D}}{\mathbf{n_D}}$ where $\mathbf{n_D}$ = the number of difference

$= \frac{\mathbf{3}}{\mathbf{6}}$ scores, which is equal to the

$= 0.50$ number of *pairs* of scores

It is important to note that while there are two sets of mileage data there is only one sample of vehicles and thus only one set of difference scores (D).

In our example, the null hypothesis is that the fuel additive does not have an effect. In other words, the null hypothesis is that there is no difference between the population means, and thus μ_D is equal to 0. The equation to determine t therefore becomes:

$$t = \frac{\mathbf{M_D - \mu_D}}{\mathbf{s_{M_D}}} = \frac{\mathbf{M_D - 0}}{\mathbf{s_{M_D}}} = \frac{\mathbf{M_D}}{\mathbf{s_{M_D}}}$$

The numerator of this equation, M_D, is simply $\sum D$ / $\mathbf{n_D}$, where $\mathbf{n_D}$ is equal to the number of difference scores, which is equal to the number of *pairs of scores*. As is indicated in Table 10.6, M_D for our example equals 3 / 6 or 0.5. It is important to recognize that this positive value of 0.5 indicates mileage is *higher* with an additive. The question we now need to address is whether this change of 0.5 miles per gallon is statistically significant, and thus indicative of a reliable effect, or whether it should simply be considered to be the result of chance.

To find the standard error, $\mathbf{s_{M_D}}$, we note that $\mathbf{s_{M_D}} = s_D / \sqrt{\mathbf{n_D}}$. And, just as the equation for the standard deviation of scores when estimating the population standard deviation (s_X) can be written as:

$$s_X = \sqrt{\frac{\mathbf{\Sigma(X - M_X)^2}}{\mathbf{n_X - 1}}}$$

s_D, the estimate of the population standard deviation of a set of difference scores (which can alternatively be defined as the estimate of the population standard deviation of the differences between pairs of scores), is equal to:

$$s_D = \sqrt{\frac{\mathbf{\Sigma(D - M_D)^2}}{\mathbf{n_D - 1}}}$$

where $\mathbf{n_D}$ is equal to the number of difference scores, and is also equal to the number of *pairs* of scores.

Substituting from Table 10.6 we have:

$$s_D = \sqrt{\frac{\mathbf{29.50}}{\mathbf{6 - 1}}}$$

$$= \sqrt{\frac{\mathbf{29.50}}{\mathbf{5}}}$$

$$= \sqrt{5.90}$$

$$= 2.43$$

We can now determine the standard error, $\mathbf{s_{M_D}}$, by noting that:

$$\begin{aligned} \mathbf{s_{M_D}} &= \frac{\mathbf{s_D}}{\sqrt{\mathbf{n_D}}} \\ &= \frac{\mathbf{2.43}}{\sqrt{\mathbf{6}}} \\ &= \frac{\mathbf{2.43}}{\mathbf{2.45}} \\ &= 0.99 \end{aligned}$$

This is the denominator that we were seeking.

The equation for t therefore becomes:

$$\begin{aligned} t &= \frac{\mathbf{M_D}}{\mathbf{s_{M_D}}} \\ &= \frac{\mathbf{0.50}}{\mathbf{0.99}} \\ &= 0.51 \end{aligned}$$

The df are $\mathbf{n_D}$ – l, where $\mathbf{n_D}$ is the number of difference scores. We therefore have 6 – 1 or 5 degrees of freedom.

The critical value from the t table for a two-tailed test with α equal to .05 and 5 df is 2.57 (Appendix K, Table 3b). Recall that as this is a two-tailed test the critical values are thus –2.57 and +2.57. As our obtained value for t is equal to +0.51, which is fewer standard deviation units from the mean than +2.57, we retain the null hypothesis and conclude that there is not enough evidence to support the view that the fuel additive changed the gas mileage of the vehicles tested. If the null hypothesis had been rejected, we would have then calculated eta squared (η^2) to indicate the effect size, where $\eta^2 = t^2 / (t^2 + df)$.

Second Example Of A Dependent Samples t Test

It was stated earlier in this chapter that while repeated measures is the most commonly used design with the dependent samples t test, you can also use this test with the matched samples design as well. For instance, let us assume that you continued to be interested in achieving better fuel economy. This time, instead of trying a fuel additive, you decide to test what effect appropriate vehicle maintenance would have. Your null hypothesis, based upon claims of advertisements, is that recommended maintenance increases fuel economy by 1 mile per gallon. Your alternative hypothesis is that it does not increase the gas mileage by this amount.

It is important to note that you need to be careful interpreting this null hypothesis. As it includes the word 'increases' you might assume that this is a one-tailed test. However, this is not the case. The word 'increases' refers to a specific standard, 1 mile per gallon, that the outcome will be compared against. The null hypothesis would be rejected if either the outcome is significantly higher or lower than this standard. Therefore, this is still a two-tailed test.

We begin by selecting pairs of different types of vehicles (large sedans, small sedans, SUVs, minivans, sports cars, etc.), all of the same approximate age. A member of each pair of vehicles is then randomly assigned to each of two groups. One group of vehicles serves as the control group. Each vehicle of the other group, the experimental vehicles, receives a tune-up. Then the gas mileage of each vehicle is determined, as is shown in Table 10.7. As usual, we will set $\alpha = .05$.

Table 10.7 Example 2: Matched Samples and Initial Calculations

Vehicle Pair	Mileage of Exp Vehicles	Mileage of Control Vehicles	Difference Scores D	$(D - M_D)$	$(D - M_D)^2$
1	16	12	4	2	4
2	15	13	2	0	0
3	14	15	–1	–3	9
4	19	17	2	0	0
5	24	20	4	2	4
6	25	25	0	–2	4
7	27	26	1	–1	1
8	30	28	2	0	0
9	33	29	4	2	4
			$\sum D = 18$	$\sum(D - M_D) = 0$	$\sum(D - M_D)^2 = 26$

$$M_D = \frac{\sum \mathbf{D}}{\mathbf{n_D}} = \frac{\mathbf{18}}{\mathbf{9}} = 2.00$$

where $\mathbf{n_D}$ = the number of difference scores, which is equal to the number of *pairs* of scores

As with the repeated measures design, with the matched samples design we begin with two sets of data but end with one set of difference scores. To test whether the null hypothesis should be rejected we once again conduct the dependent samples t test. The current null hypothesis is that the vehicles with a tune-up will have a 1 mile per gallon greater fuel economy than the vehicles without a tune-up, and thus μ_D, which reflects the hypothesized effect of the experimental treatment, is equal to +1. Since μ_D is not equal to 0 we must use the more complete version of the dependent t test equation:

$$t = \frac{\mathbf{M_D - \mu_D}}{\mathbf{s_{M_D}}}$$

where the standard error, $\mathbf{s_{M_D}} = s_D / \sqrt{\mathbf{n_D}}$

There are two terms in the numerator of the t equation, and since we have a value for each you might reasonably assume that all we have to do is to simply substitute them into the numerator. However, using a dependent t test requires understanding as well as the ability to calculate. We have indicated that the null hypothesis is that fuel economy will increase by 1 mile per gallon with proper maintenance. Thus, μ_D is equal to +1. In other words, increased fuel economy is associated with an increase in scores. Inspection of Table 10.7 also indicates that an increase in fuel economy in the experimental vehicles is associated with positive values for D. Thus, we have been consistent, as an improvement in mileage corresponds to a positive value for M_D, not a decrease. However, whether M_D is positive or negative simply reflects the order in which the treatment conditions were listed in the table. Thus the researcher has to be careful that they have been consistent. We have been, so M_D is entered into the equation for t as 2, not –2. The numerator is, therefore, 2 – 1:

$$t = \frac{\mathbf{M_D - \mu_D}}{\mathbf{s_{M_D}}}$$

$$= \frac{\mathbf{2 - 1}}{\mathbf{s_{M_D}}}$$

The denominator, $\mathbf{s_{M_D}}$, is equal to s_D / $\sqrt{\mathbf{n_D}}$. To determine the standard deviation, s_D, we once again utilize the following equation:

$$s_D = \sqrt{\frac{\mathbf{\Sigma(D - M_D)^2}}{\mathbf{n_D - 1}}}$$

where $\mathbf{n_D}$ is equal to the number of difference scores, and is also equal to the number of *pairs* of scores.

Substituting from Table 10.7 we have

$$s_D = \sqrt{\frac{\mathbf{26}}{\mathbf{9 - 1}}}$$

$$= \sqrt{\frac{\mathbf{26}}{\mathbf{8}}}$$

$$= \sqrt{3.25}$$

$$= 1.80$$

We can now determine the standard error of the mean difference ($\mathbf{s_{M_D}}$) by noting that:

$$\mathbf{s_{M_D}} = \frac{\mathbf{s_D}}{\sqrt{\mathbf{n_D}}}$$

$$= \frac{\mathbf{1.80}}{\sqrt{\mathbf{9}}}$$

$$= \frac{\mathbf{1.80}}{\mathbf{3}}$$

$$= 0.60$$

This is the denominator of the equation for t.

The calculation of t therefore becomes:

$$t = \frac{\mathbf{M_D - \mu_D}}{\mathbf{s_{M_D}}}$$

$$= \frac{\mathbf{2 - 1}}{\mathbf{0.60}}$$

$$= \frac{\mathbf{1}}{\mathbf{0.60}}$$

$$= 1.67$$

The df are $\mathbf{n_D}$ – l where $\mathbf{n_D}$ is the number of difference scores. We therefore have 9 – 1 or 8 df.

The critical value from the t table for a two-tailed test with $\alpha = .05$ and 8 df is 2.31 (Appendix K, Table 3b). As this is a two-tailed test the critical values are thus –2.31 and +2.31. Since our obtained t is equal to +1.67, which is less than +2.31, we retain the null hypothesis and conclude that there is not enough evidence to support the claim that the maintenance changed the gas mileage of the vehicles tested by other than 1 mile per gallon. However, we would recognize that the sample size is very small.

In order to illustrate how the specification of the null hypothesis can affect the outcome of a study, let us assume that the original null hypothesis had been that vehicle maintenance does not affect fuel economy. *(Note that this example is solely for illustration purposes. In a research situation you cannot re-state your null hypothesis once you have started to collect data. To do so would be unethical.)*

In this case, the hypothesized difference between the population means for the vehicle fuel economies would have been zero. We could, therefore, use the shorter version of the dependent t equation, or simply substitute 0 for μ_D in the numerator of the longer version:

$$t = \frac{\mathbf{M_D - \mu_D}}{\mathbf{s_{M_D}}}$$

The numerator now becomes 2 – 0 which is equal to 2. (Remember, we are using a positive number to indicate an increase in fuel economy.) As none of the scores have changed, the denominator remains unchanged. The t equation thus becomes:

$$t = \frac{\mathbf{2.00}}{\mathbf{0.60}}$$

$$= 3.33$$

The df remain $\mathbf{n_D}$ – l where $\mathbf{n_D}$ is the number of difference scores. We therefore continue to have 9 – 1 or 8 df.

We found previously that the critical value from the t table for a two-tailed test with $\alpha = .05$ and 8 df is 2.31. As this is a two-tailed test the critical values are thus –2.31 and +2.31. Our obtained t is equal to +3.33, which is greater than +2.31. Thus, we would now reject the null hypothesis that maintenance does not affect mileage and conclude that the maintenance changed the gas mileage of the vehicles tested. Clearly, how the null hypothesis is stated matters!

If this had been our original null hypothesis, we would proceed by calculating eta squared to indicate the effect size:

$$\eta^2 = \frac{\mathbf{t^2}}{\mathbf{t^2 + df}}$$

where η^2 is the percent of variance explained by the treatment.

For our example with 8 df, this would be:

$$\eta^2 = \frac{\mathbf{3.33^2}}{\mathbf{3.33^2 + 8}}$$

$$= \frac{\mathbf{11.09}}{\mathbf{11.09 + 8}}$$

$$= \frac{\mathbf{11.09}}{\mathbf{19.09}}$$

$$= .58 \text{ or } 58\%$$

Thus, in this example the treatment would have accounted for 58% of the total variance.

Confidence Interval for an Experiment that could have utilized the Dependent Samples t Test

If, instead, a researcher was interested in estimating the population value for the change in miles per gallon for vehicles with proper maintenance they would calculate a confidence interval. The procedure for finding a confidence interval for a dependent samples design is almost identical to what was used with a design appropriate for a one-sample t:

For one-sample t:

$$M - t_c\,(s_M) \leq \mu \leq M + t_c\,(s_M)$$

For dependent samples t:

$$M_D - t_c\,(\mathbf{s_{M_D}}) \leq \mu_D \leq M_D + t_c\,(\mathbf{s_{M_D}})$$

For our just completed example we would have:

$$2.00 - (2.31)\,(0.60) \leq \mu_D \leq 2.00 + (2.31)\,(0.60)$$

this equals:

$$2.00 - 1.39 \leq \mu_D \leq 2.00 + 1.39$$

$$0.61 \leq \mu_D \leq 3.39$$

In other words, with 8 df there is a 95% probability that a confidence interval with values between 0.61 and 3.39 miles per gallon increase in fuel economy will include the experimental population mean. (However, refer to the clarification included with the discussion of the confidence interval when using z.)

Reporting The Results Of A Dependent Samples t Test

If the original null hypothesis had been that maintenance does not affect mileage, then in an article we would report that performing maintenance on a vehicle resulted in a statistically significant increase in the gas mileage ($M = 2.00$, $SD = 1.80$). We could report that $t(8) = 3.33$, $p < .05$, $\eta^2 = .58$). However, now we can use statistical packages to specify the p-value, and it is expected that the confidence interval would also be included. Using SPSS we would say that $t(8) = 3.33$, $p = .010$, $\eta^2 = .58$, *CI* [0.61, 3.39]. Note that the p-value of .010 is less than our α of .05, confirming that we would reject the null hypothesis, and that the other values are the same as we obtained in our calculations.

Purpose And Limitations Of Using The Dependent Samples t Test

1. *Test for difference.* The null hypothesis is usually that the treatment does not have an effect. Thus if the null is correct any difference between the treatment condition means is due to chance. The alternative hypothesis is that the treatment does have an effect and the difference is not due to chance. The dependent samples t test is employed to differentiate between these two hypotheses.
2. *Does not provide a measure of effect size*. The dependent samples t test is a test of significance. It indicates whether or not an outcome is likely to have occurred by chance if the null hypothesis is correct. If the t test is significant, a measure of effect size, such as eta squared, should then be calculated.
3. *Compares a difference mean to a hypothetical difference.* With the repeated measures design the difference being analyzed is obtained from two measures from the same subject. With the matched samples design the difference is obtained from two subjects paired on some important variable. In both cases the dependent samples t test compares this obtained difference with the difference specified in the null hypothesis.
4. *Carryover effects are a concern.* A repeated measures design is a type of **longitudinal study**. In a longitudinal study subjects are measured repeatedly across time. A concern with any longitudinal study is that the effect of a treatment or intervention at one point in time may have an effect or carry over to another point in time. For instance, for most of us running 5 miles in the morning is likely to affect how quickly we can climb stairs in the afternoon. One solution to control for **carryover effects** is to employ **counterbalancingcoun**. In counterbalancing, half of the subjects are exposed to condition A first, and then later to condition B. The other half of the subjects are first exposed to condition B, and subsequently to condition A. With our example, half of us would run in the morning and climb stairs in the afternoon. The other half would climb stairs in the morning and run in the afternoon. You should note that counterbalancing will not always be effective. An

improvement in our example with running and climbing stairs would be to use counterbalancing but also to have a more lengthy rest period between the running and the climbing.

Longitudinal study – A study in which subjects are measured repeatedly across time. A repeated-measures design is a type of longitudinal study.

Carryover effect – A treatment or intervention at one point in time may affect or carry over to another point in time.

Counterbalancing – A method used to control for carryover effects. In counterbalancing, the order of the treatments or interventions is balanced so that an equal number of subjects will experience each order of presentation.

Assumptions Of The Dependent Samples t Test

1. *Interval or ratio data.* The data are on either an interval or a ratio scale of measurement.
2. *Random sample(s).* The sample in a repeated measures design is drawn at random from a population. The samples in a matched samples design are determined by randomly assigning a member of each pair to each of the two conditions.
3. *Data within each treatment level are independent.* The datum from one subject is not affecting the datum from another.
4. *Normal distribution.* The population of difference scores (D) (refer to Tables 10.6 and 10.7) is normally distributed. However, as stated in the Central Limit Theorem, the probabilities in the t table will be accurate so long as the sample size is at least 30. If the sample size is less than 30, then it is important that the underlying population be normally distributed. If you cannot collect a larger sample and do not know if the assumption of normality has been met, it would be best to turn to an alternative test.

Effect Of Violating The Assumptions

The assumption that is most likely to be violated is that the underlying population of difference scores is normally distributed. However, the t test continues to lead to accurate decisions even when this assumption is violated so long as the sample size is at least 30. If the sample size is small and you are unsure the population is normal, then you should not use the dependent samples t test. Instead, you should consider converting your interval or ratio data into an ordinal measurement and then turn to an appropriate test for the same experimental design.

Conclusion

The dependent samples t test is much like the one-sample t test. Both assume that the samples are drawn from normally distributed populations and their research designs can each be used to test an hypothesis or to construct a confidence interval. Finally, the steps involved in calculating both t tests are very similar.

The benefit of the dependent samples t test over the independent samples t is that by utilizing repeated measures or matched samples variability is likely to be reduced and thus the resulting t statistic is likely to be larger. However, degrees of freedom are lost. And, as you will see in Chapter 12, compared to the one-way within-subjects ANOVA, the dependent samples t test is probably somewhat easier to calculate than the ANOVA but it is more limited.

Comparison Of The Dependent Samples t Test And The Independent Samples t Test

The beginning of this chapter described the independent samples t test. This was followed with a review of the dependent samples t test. While the independent samples t test is used more frequently there are definite advantages to repeated-measures or matched-samples designs which employ the dependent t test. And there are disadvantages.

There are fundamentally three disadvantages to the repeated measures and matched samples designs. First, these studies frequently entail more work to conduct. With matched-samples studies, for instance, you need to identify the variable on which to match (e.g., intelligence quotient, height, personality, age, etc.), measure each subject on this variable, form pairs of similar subjects, and then randomly assign a member of each pair to each of the two groups. These steps can be time-consuming.

Second, there is an increased risk of losing subjects. With repeated-measures studies you may need the subjects to return for a second test. This may lead to the loss of subjects. And with matched-samples designs, if one member of a pair drops out of the study, you lose the data from both.

Third, in a sense you lose half of your degrees of freedom compared to an independent samples design. For a dependent samples t test, the degrees of freedom are determined by $\mathbf{n_D} - 1$, where $\mathbf{n_D}$ is equal to the *number of difference scores, which is equal to the number of pairs of scores*. Thus you have one degree of freedom for every two scores that you collect because it takes a pair of scores to obtain one difference measure. With the independent samples t test the degrees of freedom are also determined by an n – 1, but in this case n is equal to the *number of scores*. Inspection of the t table will indicate that a reduction in the number of degrees of freedom will translate into a larger critical value for t, and thus with a dependent samples t test a larger difference is required for the null hypothesis to be rejected.

Considering these drawbacks it may seem surprising that the dependent samples t test is ever used. However, there are substantial advantages to offset the disadvantages just listed. One of the major advantages of the repeated measures design is that you get more information from each subject than you would with an independent samples design. This becomes critical if you are dealing with hard-to-obtain subjects. For instance, if you were interested in the efficacy of an intervention with males between 13 and 16 years of age who are undergoing a specific form of therapy for a particular condition, your subject pool is likely to be severely limited. It makes sense, therefore, to obtain as much data as possible from each subject.

Another reason to employ a repeated measures or matched subjects design is to reduce the amount of variability in the denominator of the t equation. If the standard error (the denominator) is reduced, the value of the obtained t will increase (assuming the magnitude of the numerator stayed the same). Both repeated measures, which in a sense uses a subject as their own control, and matched samples, where a subject is paired with someone who is similar on some measure, are techniques that are likely to reduce the size of the standard error.

In effect, therefore, choice of the dependent samples t test is a balancing act. The researcher gains by increasing the amount of information obtained from each subject and by the potential to reduce the standard error which will likely lead to an increase in the size of the t ratio. But the researcher pays a price in added work, greater risk of losing subjects and the loss of degrees of freedom.

Glossary Of Terms

Carryover effect – A treatment or intervention at one point in time may affect or carry over to another point in time.

Counterbalancing – A method used to control for carryover effects. In counterbalancing, the order of the treatments or interventions is balanced so that an equal number of subjects will experience each order of presentation.

Dependent samples t test – An inferential procedure for comparing two sample means based upon repeated measures of the same subjects, or measures from pairs of subjects who are related in some way.

Difference score (D) – The difference between two measurements from the same individual (repeated measures design) or two measurements from pairs of matched subjects (matched subjects design).

Independent samples t test – An inferential procedure for comparing two means from unrelated *samples.*

Levene's test of equality of variances – Procedure used with SPSS to test the assumption that samples are drawn from populations which have equal variances.

Longitudinal study – A study in which subjects are measured repeatedly across time. A repeated-measures design is a type of longitudinal study.

Matched subjects design – A research design in which equivalent subjects are paired and then one of the subjects is randomly assigned to each group.

Repeated measures design – A research design in which each subject is tested more than once.

Standard error of the difference between sample means ($\mathbf{s}_{(M_1 - M_2)}$) – The standard deviation of the sampling distribution of the difference between sample means.

Standard error of the mean difference (s_{M_D}) – The standard deviation of the means of difference scores. More precisely, the standard deviation of the sampling distribution of the means of difference scores.

Questions – Chapter 10 - Independent Samples t

(Answers are provided in Appendix J.)

1. The independent samples t test compares _____ groups of _____ data.
   a. Two; nominal
   b. Two or more; interval/ratio
   c. Two; interval/ratio
   d. Two or more; ordinal

2. The advantage of the independent samples t test is that it is relatively easy to calculate. However, the one-way between-subjects ANOVA is _____.
   a. Even easier to calculate
   b. More flexible
   c. Able to deal with ordinal data
   d. None of the above

3. The logic of the independent samples t test directly parallels the logic of the one sample t test, for a measure of difference found in the numerator is converted into _____ by dividing by a standard error.
   a. Standard deviation units
   b. An F ratio
   c. A correlation
   d. A measure of effect size

4. Following a significant t test we would calculate _____.
   a. Post hoc tests
   b. eta squared (η^2)
   c. a linear regression
   d. we don't calculate anything, we're finished

5. All else being equal, the means of larger samples would be expected to vary _____ the means of smaller samples.

a. More than
b. The same as
c. Less than

For questions 6 to 19 use the following information: An experimenter is interested in whether drinking warm milk before going to bed will change how people sleep. The experimenter randomly assigns each subject to one of two groups. The data consist of subjects' ratings of how well they slept, with higher ratings indicating better sleep:

Control group: 3, 4, 6, 5, 2
Experimental group: 5, 9, 6, 8

6. What is the null hypothesis?
a. Drinking warm milk will change how people sleep.
b. Drinking warm milk will not change how people sleep.

7. What is the alternative hypothesis?
a. Drinking warm milk will change how people sleep.
b. Drinking warm milk will not change how people sleep.

8. Is this a one- or two-tailed test?
a. One-tailed
b. Two-tailed

9. What is the mean of the control group?
a. 2.00
b. 3.50
c. 4.00
d. 6.67

10. What is the mean of the experimental group?
a. 3.60
b. 7.00
c. 8.00
d. 8.20

11. What is the variance (s_X^2) of the control group?
a. 1.13
b. 2.50
c. 3.33
d. 4.67

12. What is the variance (s_X^2) of the experimental group?
a. 1.13
b. 2.50
c. 3.33
d. 4.67

13. What is the standard error of the difference between means?
a. 1.13
b. 2.50
c. 3.33
d. 4.67

14. What is the value of t? (Ignore the sign if your value is the same as an answer below, but negative. This would simply indicate that the order, which is arbitrary, of the experimental and control group means was reversed in the t equation.)
    a. 2.65
    b. 2.95
    c. 3.14
    d. 5.20

15. How many degrees of freedom are there?
    a. 9
    b. 7
    c. 5
    d. 3

16. What is the critical value of t from the table with alpha set to .05?
    a. 3.67
    b. 1.46
    c. 1.99
    d. 2.37

17. The outcome is _____.
    a. Statistically significant
    b. Not statistically significant

18. The value of eta squared (η^2) is _____.
    a. .00
    b. .23
    c. .36
    d. .50

19. The 95% confidence interval would be from _____ to _____.
    a. 4.00; 7.00
    b. 2.50; 3.33
    c. 0.33; 5.67
    d. 1.13; 2.65

Questions – Chapter 10 - Dependent Samples t

20. The dependent samples t test is a procedure that a researcher can use to increase the magnitude of t by reducing the variability, thereby reducing the size of the _____.
    a. Numerator
    b. Denominator
    c. Difference between means

21. If pairs of similar subjects are chosen, and then one member of each pair is randomly assigned to each condition, this is called a (an) _____ design.
    a. Repeated measures
    b. Independent samples
    c. Matched-samples
    d. Inappropriate

22. Compared to an independent samples design, with repeated measures you are gathering _____ information from each subject.
    a. More
    b. Less
    c. Exactly the same amount of
    d. Approximately the same amount of

23. If the dependent samples t test is found to be significant, we then _____.
    a. Conduct Tukey HSD tests
    b. Calculate eta squared (η^2) to indicate the effect size
    c. Use a chi-square test to determine where the significance is
    d. Stop – we are finished

24. There are disadvantages to the repeated–measures and matched-samples designs. Which of the following is <u>not</u> a disadvantage?
    a. They frequently entail more work.
    b. They can be time consuming.
    c. There is an increased risk of losing subjects.
    d. You lose degrees of freedom
    e. All of the above are disadvantages

25. A concern(s) with any longitudinal study is (are) _____.
    a. That the study will always take many years to complete
    b. Carryover effects
    c. Need to employ additional experimenters
    d. All of the above

For questions 26 – 35 use the following information: A faculty member wishes to determine whether exercise will influence the number of classes statistics students miss. There are four subjects. In the control condition the students *do not* exercise, while in the experimental condition they *do* exercise. The following data indicate the number of classes missed in each condition:

Subject	Control Condition	Experimental Condition
1	6	5
2	4	3
3	2	4
4	7	8

26. What is the null hypothesis?
    a. Exercise has no effect
    b. Exercise has an effect

27. What is the alternative hypothesis
    a. Exercise has no effect
    b. Exercise has an effect

28. Is this a one- or two-tailed test?
    a. One-tailed
    b. Two-tailed

29. What is the mean of the difference scores?

a. –2.10
b. –0.25
c. –2.70
d. 3.00

30. What is the value of s_D ?
a. 2.25
b. 0.75
c. 1.50
d. –.333

31. What is the value of s_{M_D}?
a. 2.25
b. 0.75
c. 1.50
d. –.333

32. What is the value of t?
a. 2.25
b. 0.75
c. 1.50
d. –.333

33. How many degrees of freedom are there?
a. 4
b. 3
c. 2
d. 1

34. What is the critical value of t from the table with alpha set to .05?
a. 1.89
b. 2.33
c. 3.18
d. 4.41

35. Is the outcome statistically significant?
a. Yes
b. No

Problems 36-42 utilize SPSS.

Using SPSS With The Independent Samples t Test

To Begin SPSS

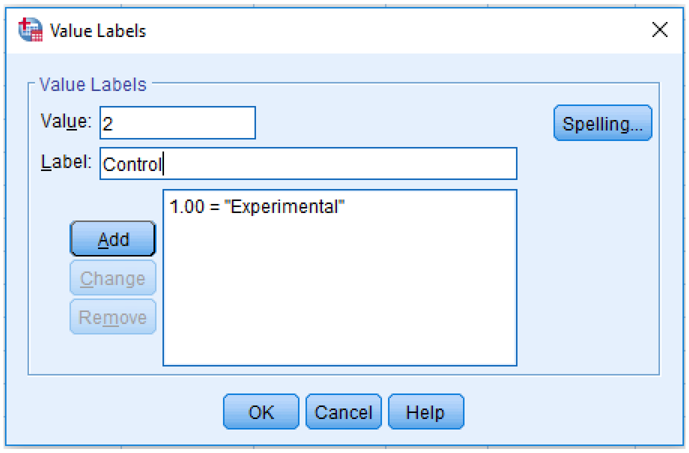
Step 1 Activate the program, close the central window, and click on the **Variable View** option at the bottom left of the window.

Step 2 Click on the first empty rectangle (called a 'cell') under the column heading 'Name' and type the name of the first variable for which you have data. We are going to utilize the same data and labels as were previously employed in Table 10.3. These data dealt with a fictitious comparison of the effectiveness of two methods for teaching statistics. We have called these procedures 'Experimental' and 'Control'. Therefore, type 'Procedure' in the first empty cell under 'Name'.

Step 3 Click on the first empty 'cell' under the column heading 'Label' and type 'Procedure of the Study'. Note that in order to see the entire label you may need to expand the size of this cell by placing your cursor on the right border of the Label heading and moving to the right.

Step 4 Click on the first empty 'cell' under the column heading 'Values'. A box will appear. In the blank space to the right of 'Value', type the number '1'. Then type a brief description of this value of the variable in the blank space to the right of 'Label'. In our case type 'Experimental'. Finally, click on 'Add'. Your label for a value of 1 will appear in the large white region in the center of the window. Now repeat the initial steps in this section for the value '2', which is given the label 'Control' (Figure 10.1). Click 'Add' and then click on 'OK'.

Figure 10.1 The Value Labels Window



Value Labels

Value Labels

Value: 2

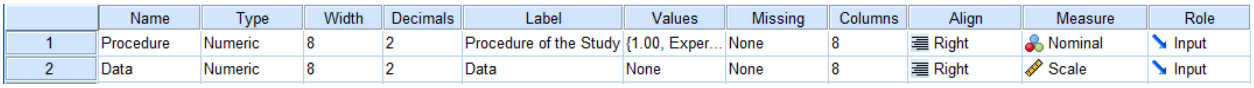
Label: Control

1.00 = "Experimental"

Step 5 Click on the first empty 'cell' under the column heading 'Measure'. As we are dealing with labels for groups, select 'Nominal'.

Step 6 Repeat Steps 2, 3 and 5 except that you type 'Data' in the first empty cell under 'Name' and for the label. Finally, select 'Scale' in the column under the column heading 'Measure' as we have ratio data. The result is shown in Figure 10.2. We must now shift to the data window and sequentially enter the data for each subject.

Figure 10.2 The Variable View Window



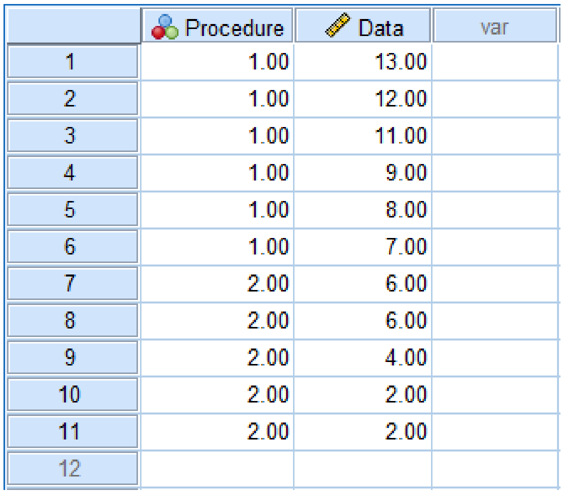
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Procedure	Numeric	8	2	Procedure of the Study	{1.00, Exper...	None	8	Right	Nominal	Input
2	Data	Numeric	8	2	Data	None	None	8	Right	Scale	Input

To Enter Data In SPSS

Step 7 Click on the 'Data View' option at the lower left corner of the variable view window.

Step 8 For each subject in the experimental condition, type the value '1' in the column 'Procedure' and their test score in the column 'Data'. Continue by entering '2' for each subject in the control condition. Then enter each subject's data (Figure 10.3).

Figure 10.3 Completed Data Entry



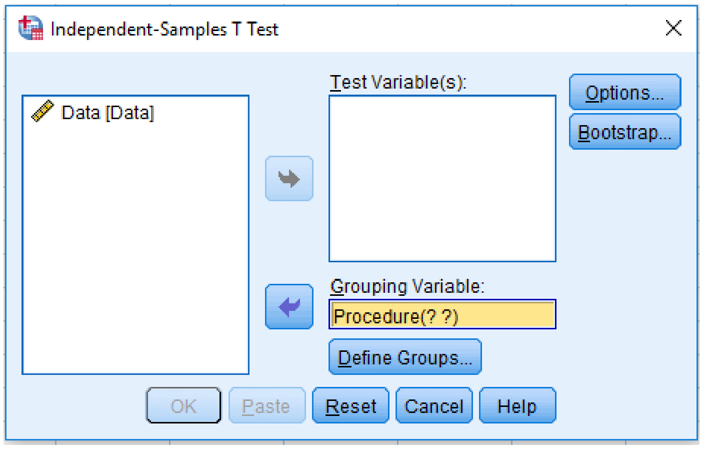
	Procedure	Data	var
1	1.00	13.00	
2	1.00	12.00	
3	1.00	11.00	
4	1.00	9.00	
5	1.00	8.00	
6	1.00	7.00	
7	2.00	6.00	
8	2.00	6.00	
9	2.00	4.00	
10	2.00	2.00	
11	2.00	2.00	
12			

To Conduct An Independent Samples t Test

Step 9 Click the cursor on '**Analyze**' along the row of SPSS commands above the data you entered, and then move to '**Compare Means**'. Then click on '**Independent Samples T Test**'.

Step 10 A new window will appear. The test variable and the grouping variable need to be identified. In our case, Procedure is the label of the grouping variable. This is indicated by moving 'Procedure' to the box under 'Grouping Variable' by clicking on the word Procedure and then on the bottom arrow. The result is shown in Figure 10.4.

Figure 10.4 The Independent Samples t Test Window



Independent-Samples T Test

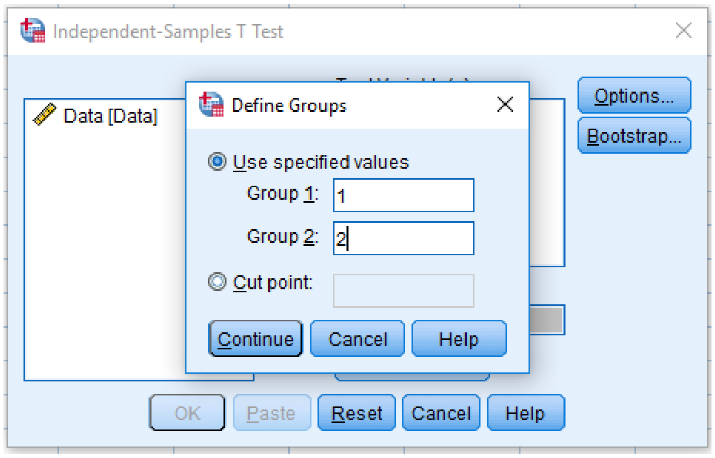
Data [Data]

Test Variable(s):

Grouping Variable: Procedure(? ?)

Step 11 Notice that there are now two question marks following the name of our grouping variable. Click on Define Groups and identify the numbers associated with our experimental and control conditions, in this case 1 and 2 (Figure 10.5). Then click '**Continue**'.

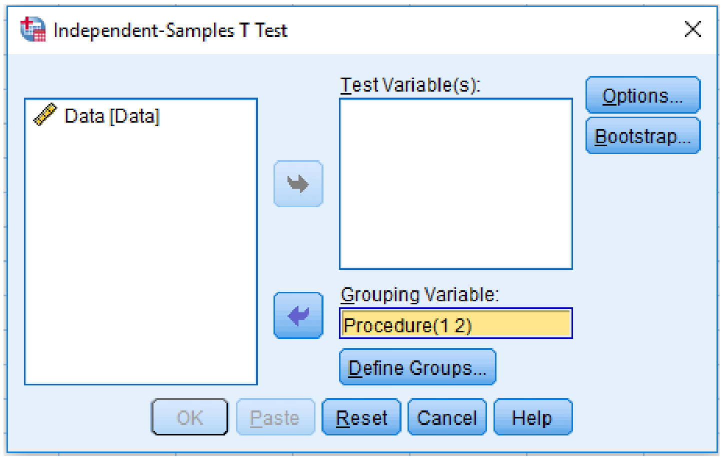
Figure 10.5 The Defining Groups Window



● Use specified values
Group 1: 1
Group 2: 2
○ Cut point:

The result will be Figure 10.6.

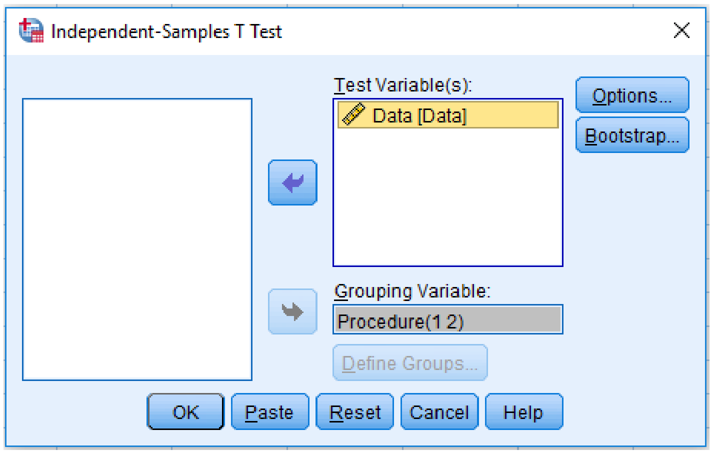
Figure 10.6 The Independent Samples T Test Window



Grouping Variable: Procedure(1 2)

Step 12 Click the word '**Data**' which will then be highlighted. Now click on the top arrow in the middle of the window. The result is shown in Figure 10.7.

Figure 10.7 Completed Independent Samples T Test Window

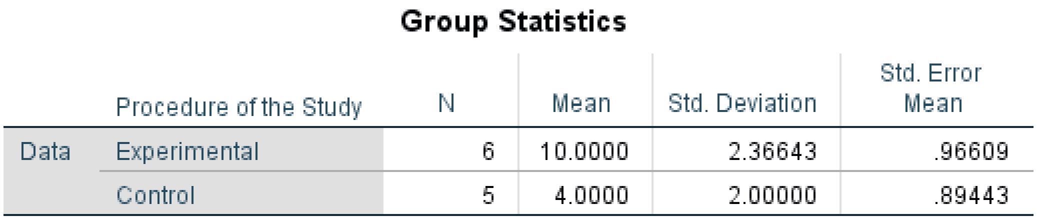


Test Variable(s): Data [Data]

Grouping Variable: Procedure(1 2)

Step 13 Now click ‘**OK**’ and SPSS will conduct the independent samples t test. The summary of the descriptive statistics are shown in Table 10.8. The values for the experimental and control group means are the same as we calculated previously. The standard deviations given in Table 10.8 are also equal to the square roots of the variances we previously calculated except for minor rounding error in our calculations.

Table 10.8 SPSS Output; Independent Samples T Test – Descriptive Statistics

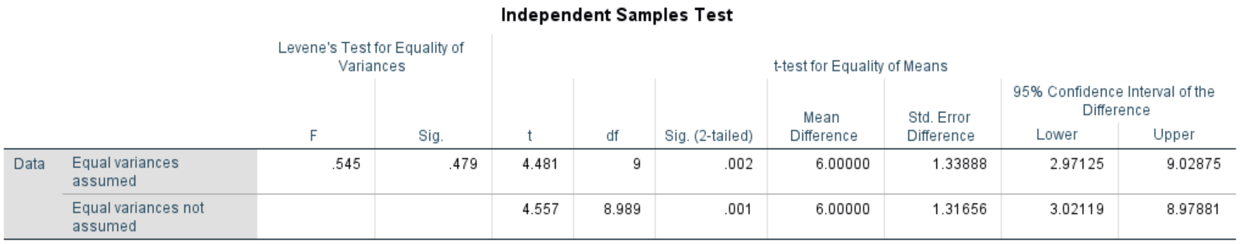


Group Statistics

	Procedure of the Study	N	Mean	Std. Deviation	Std. Error Mean
Data	Experimental	6	10.0000	2.36643	.96609
	Control	5	4.0000	2.00000	.89443

The summary of the inferential statistics are shown in Table 10.9.

Table 10.9 SPSS Output; Independent Samples T Test – Summary Table



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Data	Equal variances assumed	.545	.479	4.481	9	.002	6.00000	1.33888	2.97125	9.02875
	Equal variances not assumed			4.557	8.989	.001	6.00000	1.31656	3.02119	8.97881

This is a complex table. Refer to the first row, ‘Equal variances assumed’. The first two entries, F and Sig, refer to Levene’s test for equality of variances. An assumption of the independent samples t test is that the two samples are drawn from populations which have equal variances. If the significance (p-value) of the Levene’s test for equality of variances is less than .05 then we reject that the two samples come from populations with equal variances. If this were the case we would report the values presented in the second row of Table 10.9. In our case, the significance value for Levene's test is .479. This value is larger than .05. We can, therefore, retain the assumption that the variances of the populations are equal and proceed to the remainder of the row 'Equal variances assumed'. The next value, which is t equals 4.481, is similar to the value of 4.44 that we calculated previously except for our rounding error, and the value of 9 for the df is the same as we calculated. The value for the 'Sig (2-tailed)' (or p-value) agrees with our decision to reject the null hypothesis as .002 is less than .05. The 'Mean Difference' of 6.0 is the same as we found earlier. The 'Std Error Difference', which is 1.33888, also closely matches the value of 1.35 we found, except for some rounding error in our calculation. Finally, the values for the 95% confidence interval also closely match what we calculated.

Step 14 Exit SPSS. There is no need to save your work.

A Limitation Using SPSS With The Independent Samples t Test

In the just completed example the null hypothesis was that there was no difference between the population means from which the samples were drawn. In the second example of the independent samples t test that was described previously in the chapter the null hypothesis was that the difference between the population means was 10 (data are given in Table 10.4). The SPSS procedure that we have just reviewed cannot account for a situation where the null hypothesis is not 0. Fortunately, in the vast majority of cases the null hypothesis for an independent t test is that the difference between the population means is 0 so this is not a serious limitation.

Using SPSS With The Dependent Samples t Test

(Note that SPSS calls this the paired samples t test.)

To Begin SPSS

Step 1 Activate the program, close the central window, and click on the **Variable View** option at the bottom left of the window.

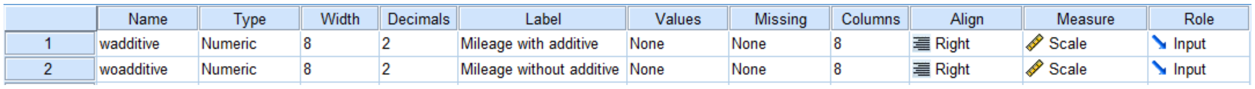
Step 2 Click on the first empty rectangle (called a 'cell') under the column heading 'Name' and type the name of the first variable for which you have data. We are going to utilize the same data as we previously employed in Table 10.6. These data dealt with a fictitious comparison of the effectiveness of a fuel additive. Click on the first empty cell under the column heading 'Name'. You now type a descriptive name of the first measure for vehicle number 1. I have chosen 'wadditive' for the mileage of a vehicle with a fuel additive.

Step 3 Click on the first empty 'cell' under the column heading 'Label'. In this cell I typed a more extensive description of the variable, 'Mileage with additive'. Note that in order to see the entire label you may need to expand the size of this cell by placing your cursor on the right border of the Label heading and moving to the right.

Step 4 Click on the first empty 'cell' under the column heading 'Measure'. As we are dealing with mileages, which are examples of ratio data, select 'Scale'.

Step 5 Repeat Steps 2 – 4 except that you type 'woadditive' in the first empty cell under 'Name' and for the label type 'Mileage without additive'. Finally, select 'Scale' in the column under the column heading 'Measure' as we have ratio data. The result is shown in Figure 10.8. We must now shift to the data window and sequentially enter the data for each subject.

Figure 10.8 The Variable View Window



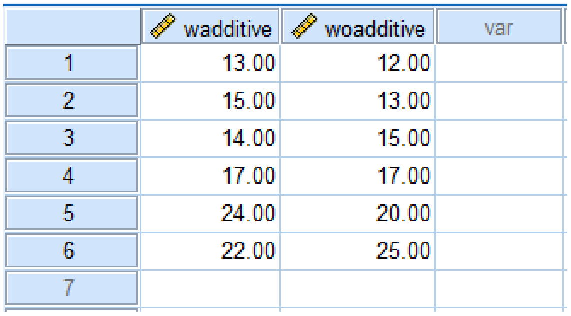
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	wadditive	Numeric	8	2	Mileage with additive	None	None	8	Right	Scale	Input
2	woadditive	Numeric	8	2	Mileage without additive	None	None	8	Right	Scale	Input

To Enter Data In SPSS

Step 6 Click on the 'Data View' option at the lower left corner of the variable view window.

Step 7 For each vehicle, type the mileage with and without an additive in the appropriate column (Figure 10.9).

Figure 10.9 Completed Data Entry



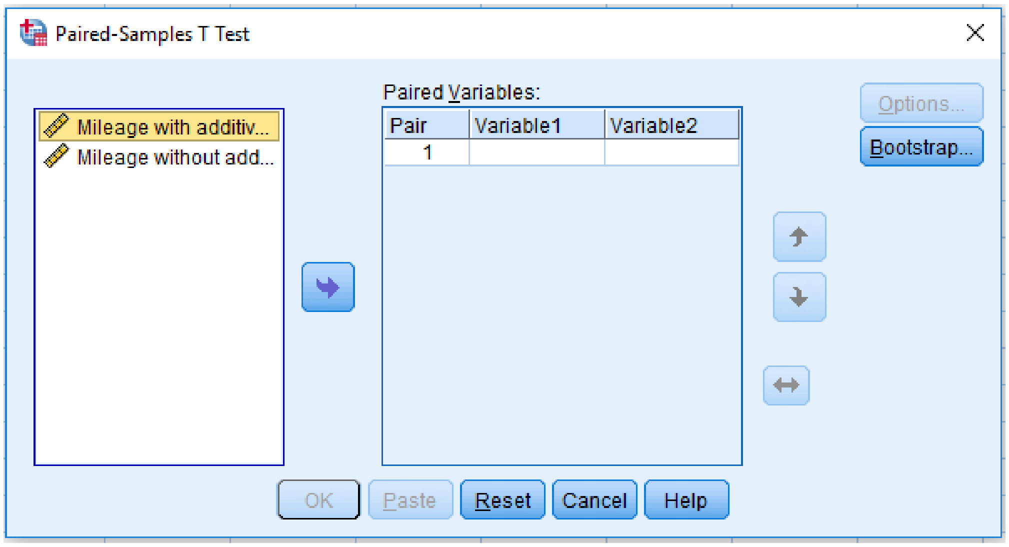
	wadditive	woadditive	var
1	13.00	12.00	
2	15.00	13.00	
3	14.00	15.00	
4	17.00	17.00	
5	24.00	20.00	
6	22.00	25.00	
7			

To Conduct A Dependent Samples t Test

Step 8 Click the cursor on '**Analyze**' along the row of SPSS commands above the data you entered, then move to '**Compare Means**'. Finally, click on '**Paired-Samples T Test**'.

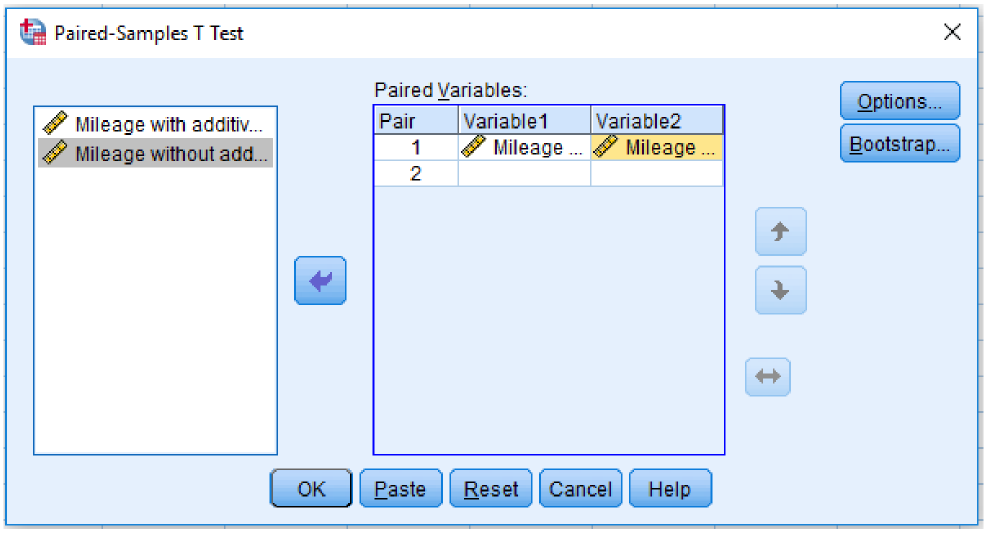
Step 9 A new window will appear (Figure 10.10).

Figure 10.10 Paired-Samples T Test Window



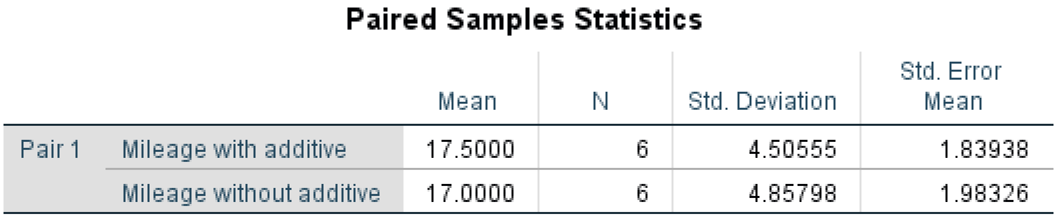
Step 10 We have only two measures for each vehicle. The first one selected will be moved to the right side of the figure under Variable 1. The second measure selected will be positioned under Variable 2. Our measure 'Mileage with additiv…' is highlighted and clicking on the arrow will copy it to the right under Variable 1. Our second measure, 'Mileage without add…' then needs to be highlighted and clicking on the arrow will copy it to the right under Variable 2. This is shown in Figure 10.11. Now click '**OK**' and SPSS will conduct a dependent samples t test.

Figure 10.11 Completed Paired-Samples t Test Window



The summary of the descriptive statistics are shown in Table 10.10. We did not previously calculate the mean and standard deviations, but these are easy to check if you would like to do so.

Table 10.10 SPSS Output; Dependent Samples t Test – Descriptive Statistics

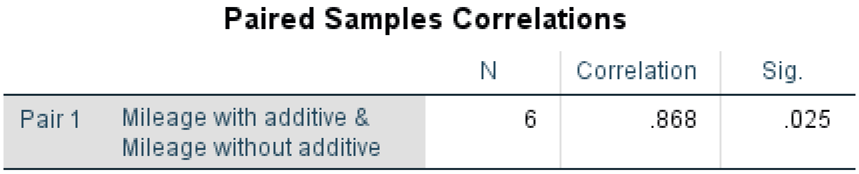


Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Mileage with additive	17.5000	6	4.50555	1.83938
	Mileage without additive	17.0000	6	4.85798	1.98326

The next portion of the output provides information about the degree to which the two measures of mileage for each vehicle are related to each other (Table 10.11). In other words, if a particular vehicle gets a high mileage without an additive does it also tend to have a high mileage when it has an additive? This topic will be covered in detail in Chapter 14 when we review correlations.

Table 10.11 SPSS Output; Relationship Between the Measures of Mileage

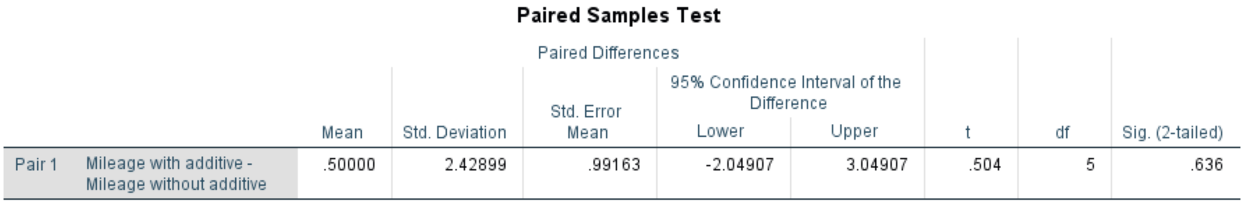


Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Mileage with additive & Mileage without additive	6	.868	.025

The summary of the remainder of the inferential statistics are shown in Table 10.12.

Table 10.12 SPSS Output; Dependent Samples t Test – Summary Table



Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Mileage with additive - Mileage without additive	.50000	2.42899	.99163	-2.04907	3.04907	.504	5	.636

There is a great deal of information given in Table 10.12. The mean difference between the mileage with and without an additive is 0.50 miles per gallon. This is the same value we obtained. Then the standard deviation of the difference in mileage is given. We calculated this value to be 2.43 which is the same as in Table 10.12 except for our minor rounding error. Further, our calculated value for the standard error was 0.99, which is essentially the same as provided in the table. SPSS then provides the lower and upper limits of the 95% confidence interval. We are also given the value of t with its df . These later values agree closely with what we calculated, and the 'Sig (2-tailed)' (or p-value) corresponds with our decision to not reject the null hypothesis as .636 is greater than our α of .05. Finally, if you compare Table 10.9, which provides the output for the Independent Samples T test, with Table 10.12, which provides the output for the Dependent

Samples T test, you will see that Levene's test is not included with the Dependent Samples T test. This is because Levene's test is only utilized when samples are independent.

Step 11 Exit SPSS. There is no need to save your work.

SPSS Problems: Chapter 10

An Example Of An Independent Samples t Test

Questions 36 – 39 are based on the following hypothetical study:

A physics professor is disturbed to learn that many students apparently do not have a good understanding of facts related to motion. For instance, many do not understand that on a moving carousel the riders on the inside are traveling more slowly than those nearer the outside edge. The professor decides to examine whether passing a college-level physics class affects students' understanding of motion. Specifically, the professor compares two groups, one consisting of students who have passed an introductory physics course and the other consisting of students who have not taken such a course. (Note that this is a quasi-experimental design.) The professor measures knowledge of motion on a 50-point scale, with higher numbers indicating better knowledge of motion. Do the data indicate that the physics course has had a statistically significant impact on student understanding? (Use a two-tailed test with alpha equal to .05.)

Scores for students who have passed a physics course	Scores for students who have not taken a physics course
45	40
47	38
41	36
38	35
46	38
44	42
42	45
42	31
	35
	47

36. What is the significance value (Sig.) for Levene's test for equality of variances?
    a. .002
    b. .083
    c. .197
    d. .359

37. What is the value of t?
    a. 1.445
    b. 2.242
    c. 3.739
    d. 4.127

38. What are the degrees of freedom?

a. 12
b. 14
c. 16
d. 18

39. Is the outcome statistically significant?
a. yes
b. no

An Example Of A Dependent Samples t Test

Questions 40 - 42 are based on the following hypothetical study:

A number of studies have indicated that conservatives are more anxious than liberals. In order to test whether an increase in anxiety changes views of conservatism/liberalism, a faculty member measures the political views of students early in a semester (Assessment 1) and again just before they take an important, cumulative final exam in a required course with a reputation for being difficult (Assessment 2). The scale goes from 0 (very liberal) to 10 (very conservative). Assuming that the following results are obtained, is there a significant shift of the political views in this hypothetical study? (Use a two-tailed test with alpha equal to .05.)

Student	Assessment 1	Assessment 2
1	6	6
2	6	7
3	4	5
4	3	4
5	5	6
6	5	7
7	9	8
8	9	10
9	6	5
10	7	6

40. What is the value of t?
a. -1.177
b. -2.361
c. -3.904
d. -4.846

41. What are the degrees of freedom?
a. 6
b. 7
c. 8
d. 9

42. Is the outcome statistically significant?
a. yes
b. no

Chapter 11
Finding Differences with Interval/Ratio Data – III: The One-way Between-subjects ANOVA

"Enter to grow in wisdom."

Inscription on the outside of the 1890 gate to Harvard Yard

Introduction

As Table 11.1 indicates, when testing for a difference with interval or ratio data a commonly used set of procedures is the **analysis of variance** (abbreviated ANOVA). It is important to recognize that the ANOVAs are a general approach to analyzing data that can be used with a variety of experimental designs.

Analysis of variance (ANOVA) – A set of flexible, closely related, inferential procedures for comparing sample means by examining variances.

In order to discuss the ANOVAs we first need to master some additional vocabulary. When using the ANOVAs an independent variable (IV) is called a **factor**, and each value of a factor (IV) is called a **level**. In other words, if we are studying the effect of hours of sleep then each different amount of sleep to which subjects are assigned is a level. ANOVAs are particularly useful because they can deal simultaneously with more than one factor (IV), and each factor can have two, or more, levels. When the design includes more than one factor we have what is called a **factorial ANOVA**. However, in this chapter we are only dealing with one factor (one IV). Thus, we will be reviewing what could be described as the single-factor ANOVA. Instead of single-factor, often the phrase 'one-way' is used. This chapter will, therefore, be introducing the one-way ANOVA. Further, when each subject is assigned to only one of the levels, this is a **between-subjects design** (this issue will be discussed further in the next chapter). Thus, the name **one-way between-subjects ANOVA** (this design is underlined in Table 11.1) conveys a great deal of information to a statistician. First, the phrase one-way identifies that we are dealing with a research design that has only one IV. Second, the phrase 'between-subjects' indicates that different subjects are assigned to each of the experimental conditions.

Factor – With an ANOVA, the term 'Factor' is often used instead of independent variable.

Level – With an ANOVA, the number of values of an independent variable.

Factorial ANOVA – An ANOVA with more than one factor.

Between-subjects design – With an ANOVA, those designs in which each subject experiences only a single level of a factor.

One-way between-subjects ANOVA – An inferential procedure for comparing two or more means from independent samples when there is one independent variable.

The ANOVAs can analyze data from both true and quasi-experimental designs. In a true experiment each subject is randomly assigned to a level of the IV. In the case of a quasi-experiment the subjects cannot be randomly assigned to the levels of the IV. For instance, if there are two levels of the IV, one consisting of men and the other consisting of women, the researcher cannot randomly assign which gender a subject will be, and thus does not control the membership of the two samples.

Table 11.1 Overview of Inferential Statistical Procedures For Finding if there is a Difference

	Type of Data			
	Nominal (Frequency)		Ordinal (Ranked)	Interval/Ratio (Continuous Measure)
Research Design		Research Design		
One Variable With At Least Two Outcomes	Goodness-of-fit Chi-Square	One IV With One Sample		One-sample z Test or One-sample t Test
		One IV With Two Or More Independent Samples	*Kruskal–Wallis H*	One-way Between–Subjects ANOVA (Only two independent samples, Independent Samples t Test)
		One IV With One Sample Having Two Or More Repeated Measures		One-way Within–Subjects ANOVA (Only two repeated measures, Dependent Samples t Test)
Two Variables, Each With At Least Two Outcomes	Chi-Square Test of Independence	Two IV Each With Two Or More Independent Samples		Two-way Between–Subjects ANOVA

The Italicized procedure is reviewed in Appendix A

The ANOVA is not the only procedure that can be used when looking for a difference between subjects when there are interval or ratio data. In Chapter 10 you learned that the independent samples t test is also commonly employed, but it can only be used with designs that have one IV with two samples of subjects. In contrast, the one-way between-subjects ANOVA can simultaneously analyze the data from studies with two, or more, values (levels) of its IV. For instance, instead of only having a control group and one experimental group, as is the case with the independent samples t test, we could now have a control group and a number of experimental groups. Thus, while the independent samples t test examines the difference between two sample means, the one-way between-subjects ANOVA provides an overall test of the significance of the differences between all of the sample means. In doing so the ANOVA controls for the **experimentwise error rate**. In other words, if α is set at .05, with an ANOVA there is only one chance in 20 of making a Type I error for the entire set of comparisons of sample means.

It is critical to understand what the last sentence indicates. When a single statistical test is calculated to determine if a difference exits between two groups, with α set to .05 there is a 5% chance of rejecting the null hypothesis when, in fact, it is correct. This is the probability of making a Type I error. As the number of comparisons is increased, with a procedure such as the independent samples t test the probability of making a Type I error *for each comparison* remains at .05. This is the **pairwise error rate**. However, the likelihood of making a Type I error *across all the comparisons* increases rapidly. Put another way, the probability of making a Type I error remains .05 for each **pairwise comparison** but the experimentwise error rate increases dramatically as the number of pairwise comparisons increases (Table 11.2). (We encountered a similar situation when discussing post hoc comparisons for the chi-square test of independence.)

Experimentwise error rate – The likelihood of making at least one Type I error with any of the experiment's comparisons.

Pairwise error rate – The likelihood of making a Type I error for a single comparison between sample means. This is equal to α, which is usually .05 or .01.

Pairwise comparison – Comparison between two sample means.

Table 11.2 Likelihood of Making at Least one Type I Error

Number of Groups or Samples	Number of Pairwise Comparisons	Pairwise Error Rate	Likelihood of at Least one Type I Error
2	1	.05	.05
6	15	.05	.54
12	66	.05	.97

As Table 11.2 indicates, as the number of groups or samples rises, the number of pairwise comparisons increases rapidly. With two groups there is only one pairwise comparison. However, with 12 samples there would be 66 pairwise comparisons. This raises two concerns. First, while calculating a single independent samples t test is not difficult, calculation of 66 t tests is definitely a chore. Second, and more important, the likelihood of making a Type I error increases dramatically when numerous t tests are calculated. It follows that the null hypothesis, which states that all of the sample means are equal, will almost certainly be rejected if a large series of t tests is conducted (Table 11.2). Clearly, what is needed is a method to control the experimentwise error rate as the number of pairwise comparisons increases, and the ANOVA accomplishes this elegantly.

Logic of a One-Way Between-Subjects ANOVA

As the name analysis of variance implies, the ANOVA is based upon a comparison of variances. It may at first seem peculiar to test differences between means by examining variances, but this has turned out to be a very useful approach. Fortunately, the logic of the ANOVA is straightforward. Specifically, *the between-subjects ANOVA is examining whether the variability observed among the scores within the treatment groups is sufficient to account for the variability observed among the means of these groups.* This can be made clearer with an example. Let's assume we take two random samples from a population. We record the height of each subject and then calculate the mean height and variance of each sample. Since both samples are randomly drawn from the same population, we would expect these two sample means to be similar and the two variances to also be similar.

Now let us assume that we add a treatment effect to one of our groups. In this case, the control group's mean and standard deviation would not change as we have not influenced them in any way. However, if the intervention for each member of the treatment group was to stand on a chair 2 feet tall, the mean height of this group would now increase by 2 feet. Since adding a constant to every score does not change the standard deviation or variance of a sample, (this was discussed in Chapter 3) the variance (s_X^2) of the treatment group would not change and would, therefore, remain approximately the same as the control group's. In other words, the addition of a treatment effect does not change the variability *within* a group. However, the addition of this treatment effect would cause the sample means to diverge since the mean of the treatment group has increased by 2 feet. Thus the variability of treatment means will have increased but the variability of scores within each of the groups will have remained unchanged. Consequently, the variability of the scores *within* the groups would no longer be sufficient to account for the variability observed *between* the group means. And, as you will see, the ANOVA will detect this inequality which is indicative of a treatment effect occurring (in this case standing on a chair). We now turn to an in-depth explanation.

The word 'analysis' can be defined as the examination of the parts that make up some whole. In an ANOVA it is the variance that will be analyzed. A one-way ANOVA starts with the assumption that the **treatment** (IV) does not have an effect and, consequently, the different groups are essentially random samples from a single population. The ANOVA then estimates the variance of this population in two different ways. Fortunately, you are familiar with both approaches.

Treatment – With ANOVA, another term for the independent variable.

We learned in Chapter 9 that if we randomly select a sample from a population the standard deviation of the sample (s_X) can be used to estimate the population standard deviation (σ_X) where:

$$s_X = \sqrt{\frac{\Sigma(X - M)^2}{n - 1}}$$

It follows, therefore, that the variance derived from a sample (s_X^2) can be used to estimate the population variance (σ_X^2) where :

$$s_X^2 = \frac{\Sigma(X - M)^2}{n - 1}$$

And if we continue to randomly select samples from this population we can combine or pool the variance estimate from each of these samples to find an even more precise estimate of σ_X^2. Stated differently, in calculating an ANOVA *one estimate of the population variance (σ_X^2) comes from looking at the variability within each of the samples*. If the scores within each sample do not vary substantially it suggests that σ_X^2 is small. In contrast, if the scores within each sample do vary substantially it suggests that σ_X^2 is large. This estimate of σ_X^2 is called the **mean square within** (MS_W).

Mean square within (MS_W) – The estimate of the population variance (σ_X^2) based upon the variability within each of the samples. More specifically, it is obtained by pooling the variances of the scores within each of the samples.

Alternatively, we could estimate the population variability by examining how much the sample means vary from each other. If the sample means do not vary substantially it suggests that the population variance (σ_X^2) is small (you would also have to account for the size of the sample). However, if the sample means do vary substantially it suggests that σ_X^2 is large. This estimate of σ_X^2 is called the **mean square between** (MS_{Bet}). (Please note that we are not calling this MS_B. The term MS_B is used in more complex ANOVAs and has a different definition than MS_{Bet}.)

Mean square between (MS_{Bet}) – The estimate of the population variance (σ_X^2) based upon the variability between the sample means. More specifically, it is obtained from the deviations of the sample means from the grand mean.

What is essential to note is that we now have two methods for estimating σ_X^2. We can find one estimate by pooling the *variability of scores within each of the samples*, and we can find another estimate based upon the *variability of the sample means*. These two estimates of σ_X^2 may, or may not, be similar.

The key to understanding ANOVA is that if the samples are all drawn from the same population these two estimates of σ_X^2 are expected to be approximately the same. However, if the samples are drawn from populations with different means the two estimates of σ_X^2 may differ substantially. An example will clarify the reasoning.

Let us assume that from a population we randomly select a sample of 30 subjects and obtain a variance (s_X^2) of 2.5. This is an indication of how much the *sample scores* vary from each other. And, as we have seen, s_X^2 provides us with an estimate of σ_X^2. Of course, if we were to draw another sample from the same population we would expect the variance of this sample to also be approximately 2.5. However, we would not be surprised if the variance differed somewhat from 2.5. And this second sample's s_X^2 would also provide us with an estimate of σ_X^2. An even better estimate of σ_X^2 would be determined by combining the s_X^2 of the first sample with the s_X^2 of the second sample. As we just reviewed, this results in what is called the mean square within (MS_W). It is important to note that the MS_W is based upon the variability *within* each sample. And since each subject within a sample is treated the same (receives the same treatment level of the independent variable) then the MS_W does not reflect any effect of the treatment. Since this variability is not due to treatment, it is called **error**. But realize that this does not signify that any mistake occurred. It is simply the variability among the subjects that is not due to the independent variable.

<u>Error</u> – With ANOVA, the variability not due to treatment.

Alternatively, as we previously noted, the population variance (σ_X^2) could also be estimated by examining how much the *sample means* vary. Specifically, we have learned in Chapter 9 that s_M (a measure of the variability of sample means) equals $s_X/\sqrt{n}$. It follows, therefore, that s_M^2 equals s_X^2/n. Thus from the sample means we can once again determine a value for s_X^2 which can be used as an estimate of σ_X^2. And we just noted that this second estimate of σ_X^2 is called the mean square between (MS_{Bet}). This estimate of σ_X^2 is affected by two sources of variability. As was the case for MS_W, the magnitude of MS_{Bet} is affected by the amount of variability, called error, within each of the samples. The larger this variability, the more we would expect the sample means to diverge just by chance. In addition, since the magnitudes of the sample means reflect the effect of the independent variable (treatment), MS_{Bet} (unlike MS_W) is also affected by the treatment. Thus MS_{Bet} is affected by both the treatment and the error, while MS_W only reflects the amount of error.

As was stated previously, it can be shown mathematically that *if there is no treatment effect, these two estimates of the population variability (σ_X^2) are expected to be approximately equal. In* *other words, if there is no treatment effect then both the mean square within (MS_W) estimate of $\sigma_X{}^2$ and the mean square between (MS_{Bet}) estimate are based solely upon what we are calling error, and they are expected to be approximately equal.* Put another way, the ratio of (MS_{Bet}) / (MS_W) should be approximately 1 if there is no treatment effect. In other words, if MS_W is approximately equal to MS_{Bet} then the variability of the scores within the groups is sufficient to account for the variability of the group means.

The ratio of (MS_{Bet}) / (MS_W) is given a name in statistics. (As you have learned, everything in statistics seems to have a name.) It is called the F ratio in honor of Sir Ronald Fisher who made major contributions to the development of the ANOVA. It is the ratio of two estimates of the population variance:

$$\mathrm{F} = \frac{\mathbf{MS_{Bet}}}{\mathbf{MS_W}} = \frac{\textbf{between groups estimate of } \sigma_{\mathbf{X}}^2}{\textbf{within groups estimate of } \sigma_{\mathbf{X}}^2} = \frac{\textbf{estimate of } \sigma_{\mathbf{X}}^2 \textbf{ based upon treatment + error}}{\textbf{estimate of } \sigma_{\mathbf{X}}^2 \textbf{ based only upon error}}$$

As you just learned, if there is no treatment effect this ratio should be approximately 1. This is because if there is no treatment then both the numerator and denominator are solely estimates of error (variability not due to any treatment). Thus, *when there is no treatment effect the F ratio would become:*

$$\mathrm{F} = \frac{\mathbf{MS_{Bet}}}{\mathbf{MS_W}} = \frac{\textbf{between groups estimate of } \sigma_{\mathbf{X}}^2}{\textbf{within groups estimate of } \sigma_{\mathbf{X}}^2} = \frac{\textbf{estimate of } \sigma_{\mathbf{X}}^2 \textbf{ based upon } \cancel{\textbf{treatment}} \textbf{ + error}}{\textbf{estimate of } \sigma_{\mathbf{X}}^2 \textbf{ based only upon error}} = 1$$

On the other hand, if there is a treatment effect the size of the numerator would increase but the size of the denominator would not change. And thus the F ratio would be greater than 1. Therefore, in order to determine if there is a treatment effect we first have to calculate the F ratio and then refer to the appropriate table to determine if the outcome is greater than would be expected to have occurred by chance. Fortunately, while our discussion has been based upon the one-way between-subjects ANOVA, the logic remains substantially the same with more complex designs.

Conducting A One-Way Between-Subjects ANOVA

While the between-subjects ANOVA can deal with more than two groups our first example will have just two in order to simplify the computations. The data deal with a comparison of quiz grades for a control group and an experimental group. These data are indicated in Table 11.3, along with the calculation of the sample means and sum of the squared deviations from each of the means. Note that for the entries in the experimental condition a value of 3 has been added to each score in the control condition. As expected, this increases the value of the experimental sample's mean by 3, but does not affect its variability.

Table 11.3 Example 1: Initial Calculations

Control Condition (X_1)			Experimental Condition (X_2)		
X_1	$(X_1 – M_1)$	$(X_1 – M_1)^2$	X_2	$(X_2 – M_2)$	$(X_2 – M_2)^2$
7	2	4	10	2	4
6	1	1	9	1	1
5	0	0	8	0	0
4	–1	1	7	–1	1
3	–2	4	6	–2	4
$\Sigma X_1 = 25$	$\Sigma x_1 = 0$	$\Sigma x_1^2 = 10$	$\Sigma X_2 = 40$	$\Sigma x_2 = 0$	$\Sigma x_2^2 = 10$
$n = 5$			$n = 5$		
$M_1 = 25 / 5$			$M_2 = 40 / 5$		
$= 5$			$= 8$		

It is critical when using ANOVAs that you know what you are calculating and that you keep your calculations clearly defined. The nine values that must be calculated in a one-way between-subjects ANOVA are underlined in Table 11.4.

Table 11.4 Summary Table for the One-way Between-subjects ANOVA

Source of Variation	SS	df	MS	F
Between Groups	SS_{Bet}	df_{Bet}	MS_{Bet}	F ratio
Within Groups	SS_W	df_W	MS_W	
Total	SS_T	df_T		

We will begin our analysis of the quiz grades by entering the three values for SS, then calculating three values for df, two values for MS and finally one F ratio. As you will see no step is difficult. It is essential, however, that you clearly identify each item you are calculating so that you do not become confused.

Calculating The Sums Of Squares

The first values in Table 11.4 that must be filled in are the sums of squares (SS). The calculations are easiest if we begin by calculating the **sum of squares total** (SS_T). This value is found by first calculating the mean of all of our scores (to keep the computations brief, in this example there are only 10 quiz scores), which is known as the grand mean, and then determining the sum of the squared deviations of each score from this grand mean. This can be represented as:

$$SS_T = \Sigma(\mathbf{X} – \mathbf{M_G})^2$$

where M_G is the mean of all of the scores, in other words the **grand mean**. It is found using the following equation:

$$M_G = \frac{\Sigma \mathbf{X}}{\mathbf{N}}$$

where, in a one-way between-subjects ANOVA, N is the total number of subjects (note that in this calculation N is not referring to the size of a population).

Sum of squares total (SS_T) *– The sum of the squared deviations from the mean for all of the scores.*

Grand mean (M_G) *– The mean of all of the scores.*

For our example with 10 quiz scores the calculation of M_G and SS_T are shown in Table 11.5. Note that all of the scores from both groups are included.

Table 11.5 Example 1: Calculation of the Sums of Squares Total

X	$(X - M_G)$	$(X - M_G)^2$
7	0.50	0.25
6	−0.50	0.25
5	−1.50	2.25
4	−2.50	6.25
3	−3.50	12.25
10	3.50	12.25
9	2.50	6.25
8	1.50	2.25
7	0.50	0.25
6	−0.50	0.25
$\Sigma X = 65$	$\Sigma(X - M_G) = 0$	$\Sigma(X - M_G)^2 = 42.50 = SS_T$
$N = 10$		
$M_G = \frac{\mathbf{65}}{\mathbf{10}}$		
$= 6.50$		

This value of SS_T is then entered in our ANOVA summary table. It is a measure of the total variability in the data.

The **sum of squares between groups** (SS_{Bet}) is found by determining the square of the deviation of each sample mean (M) from the grand mean (M_G), then multiplying by the *sample* size (n), and finally finding the sum for all of the samples. Thus, conceptually:

$SS_{Bet} = \Sigma\,[(\mathbf{M} - \mathbf{M_G})^2\mathbf{n}]$

We have already determined that $M_G = 6.50$ (Table 11.5). From Table 11.3 we find that n, the size of *each sample* (note that $n \neq N$), is equal to 5 and that the two sample means (M) are 5 and 8. For our example the calculations for determining SS_{Bet} are shown in Table 11.6.

<u>Sum of squares between groups</u> (SS_{Bet}) – The sum of the squared deviations of each treatment mean from the grand mean.

Table 11.6 Example 1: Calculation of SS_{Bet}

	M	$(M - M_G)$	$(M - M_G)^2$	$(M - M_G)^2n$
	Becomes:	$(M - 6.50)$	$(M - 6.50)^2$	$(M - 6.50)^2(5)$
Control	5.00	–1.50	2.25	$(2.25)(5) = 11.25$
Exp	8.00	1.50	2.25	$(2.25)(5) = \underline{11.25}$
				$\Sigma(M - M_G)^2\,n = 22.50 = SS_{Bet}$

Alternatively, the same calculations can be presented as follows:

$$\begin{aligned} SS_{Bet} &= \Sigma\,[(\mathbf{M} - \mathbf{M_G})^2\mathbf{n}] \\ &= [(5 - 6.50\,)^2\,(5)] + [(8 - 6.50)^2\,(5)] \\ &= [(-1.50)^2\,(5)] + [(1.50)^2\,(5)] \\ &= (2.25)(5) + (2.25)(5) \\ &= 11.25 + 11.25 \\ &= 22.50 \end{aligned}$$

This value is then entered in our ANOVA summary table. It is a measure of the variability of the group means.

The **sum of squares within groups** (SS_W) can be found by taking each score (X), subtracting its sample mean (M), and squaring this deviation. The sum of these squared deviations for all of the scores in each of the groups would be SS_W. Thus, conceptually:

$$SS_W = \Sigma[\Sigma(X - M)^2]$$

where M is the mean of a group or sample.

In other words, with only two groups:

$$SS_W = \Sigma x_1{}^2 + \Sigma x_2{}^2$$

Fortunately, we have already calculated these values in Table 11.3. Therefore:

$$SS_W = 10 + 10$$

$= 20$

This value is then entered into our ANOVA summary table. It is a measure of the variability within each of the groups.

<u>Sum of squares within groups</u> (SS_W) – The sum across all conditions, of the sum of the squared deviations of each score from its treatment mean.

In order to check our calculations, we make use of the fact that:

$SS_T = SS_{Bet} + SS_W$

$42.50 = 22.50 + 20$

$42.50 = 42.50$

Calculating the three SS is the most challenging step in completing the ANOVA summary table. It is important that you understand what you have accomplished, and why the next steps are necessary. A SS is a measure of variability. Unfortunately neither the SS_{Bet} nor the SS_W is a completely adequate measure of variability as the magnitude of SS_{Bet} is affected by the number of groups and the magnitude of SS_W is affected by the number of subjects. By dividing each SS by the appropriate degrees of freedom we control for the number of groups and the number of subjects. The result in each case is what is called a mean square (MS). This is another term for variance. And then these two variances are directly compared in what is called an F ratio. That's all there is to calculating a one-way between-subjects ANOVA.

Calculating The Degrees Of Freedom

We now must calculate three values of degrees of freedom; the degrees of freedom for between groups, within groups, and total. For a one-way between-subjects ANOVA the degrees of freedom for between groups is equal to the number of groups minus 1. Thus:

$df_{Bet} = k – 1$

where k is the number of levels of the IV. In our example:

$df_{Bet} = 2 – 1$

$= 1$

This value is then entered in our ANOVA summary table.

To find the degrees of freedom for within groups we first subtract 1 from the total number of subjects in each group and then sum the resulting values across all of the groups. Thus:

$df_W = \Sigma(n – 1)$

where n is the number of subjects in each group or sample. In our example:

$df_W = (5 – 1) + (5 – 1)$

$= 4 + 4$

$= 8$

Alternatively, the following equation can be used to calculate df_W:

$df_W = N – k$

where N is the total number of subjects in all the groups or samples and where k is the number of groups or samples. For our example:

$df_W = 10 – 2$

$= 8$

The value for df_W is then entered in our ANOVA summary table.

To find the degrees of freedom total, we subtract 1 from the total number of subjects. Thus:

$df_T = N – 1$

where N is the total number of subjects in all the groups or samples. In our example:

$df_T = 10 – 1$

$= 9$

This value is then entered in our ANOVA summary table.

As a check on our calculations:

$df_T = df_{Bet} + df_W$

$9 = 1 + 8$

$9 = 9$

Calculating The Mean Squares

Two **mean squares** (MS) now need to be calculated. The MS between groups and the MS within groups are found by dividing the appropriate SS by its degrees of freedom. Thus:

$MS_{Bet} = \frac{\mathbf{SS_{Bet}}}{\mathbf{df_{Bet}}}$ $\qquad$ $MS_W = \frac{\mathbf{SS_W}}{\mathbf{df_W}}$

$= \frac{\mathbf{22.50}}{\mathbf{1}}$ $\qquad$ $= \frac{\mathbf{20.00}}{\mathbf{8}}$

$= 22.50$ $\qquad$ $= 2.50$

These values are then entered in our ANOVA summary table. Recall that each MS is an estimate of the population variability (remember, each is a variance). However, each estimate is derived from a different perspective, from looking at the variability *between* the group means (MS_{Bet}), and from looking at the variability *within* each of the groups (MS_W). And recall that MS_{Bet} reflects both treatment and error while MS_W only reflects error.

Mean square (MS) – In an ANOVA, an estimate of the population variance (σ_X^2).

Calculating The F Ratio

Finally, we calculate our F ratio:

$$F = \frac{\mathbf{MS_{Bet}}}{\mathbf{MS_W}}$$

$$= \frac{\mathbf{22.50}}{\mathbf{2.50}}$$

$$= 9.00$$

This value is then entered in our summary table, and the ANOVA table is complete (Table 11.7). We will later calculate the value listed in the final column of this table.

Table 11.7 Example 1: Completed Summary Table for the One-way Between-subjects ANOVA, with the Value for Eta Squared (η^2)

Source of Variation	SS	df	MS	F	η^2
Between Groups	22.50	1	22.50	9.00	.53
Within Groups	20.00	8	2.50		
Total	42.50	9			

Interpreting The F Ratio

As was described previously, MS_{Bet} and MS_W are each estimates of the population variability and they would be expected to be similar if the independent variable did not have an effect. If this were the case, we would expect the F ratio to be approximately equal to 1. However, if the independent variable had an effect this would increase the differences among the group means, which would lead to an increase in the MS_{Bet}. But MS_W would not be affected. Consequently, the F ratio would now be greater than 1.

To determine whether our calculated F ratio of 9.00 is significantly different from a value of 1, which is what would be expected if the independent variable had no effect, we must enter the F table (Appendix K, Table 4). The F ratio is based upon two MS estimates, each with its degrees of freedom. To find the critical value of F for alpha equal to .05 we locate the column corresponding to the degrees of freedom we used in the calculation of the numerator of our F ratio, and the row corresponding to the degrees of freedom we used in the calculation of the denominator of our F ratio. For our F this would be 1 and 8 degrees of freedom. At the intersection of this column and row the critical value of F is 5.32. As our obtained F of 9.00 is larger than the critical value we reject the null hypothesis that the samples came from populations with equal means and accept the alternative hypothesis that the population means differ. Put differently, we conclude that our independent variable had an effect.

Calculating The Post Hoc Comparisons

A significant F indicates that the independent variable had an effect. As our independent variable had only two levels, the effect has to be due to a difference between the means of the control and experimental groups. If there were more than two levels we would need to conduct post hoc comparisons, as we did with the chi-square test of independence, to determine which treatment level means differ.

Calculating The Effect Size

For the one-way between-subjects ANOVA we will utilize eta squared (η^2) as the measure of effect size. η^2 for this ANOVA is easily calculated by hand:

$$\eta^2 \text{ for treatment} = \frac{\mathbf{SS_{Bet}}}{\mathbf{SS_T}}$$

It is the proportion of the total variability that is explained by the treatment. In our example:

$$\eta^2 \text{ for treatment} = \frac{\mathbf{22.50}}{\mathbf{42.50}}$$

$$= .53 \text{ or } 53\%$$

This value of η^2 is included in the last column of Table 11.7.

We can also utilize η^2 to determine the proportion of the total variability that is *not* explained by the treatment. The equation for this η^2 is:

$$\eta^2 \text{ for error} = \frac{\mathbf{SS_W}}{\mathbf{SS_T}}$$

$$= \frac{\mathbf{20.00}}{\mathbf{42.50}}$$

$$= .47 \text{ or } 47\%$$

A useful characteristic of η^2 values when used with ANOVAs is that their sum will equal 1.00. For example, in our case η^2 for treatment + η^2 for error = .53 + .47 = 1.00.

Reporting The Results Of A One-Way Between-Subjects ANOVA

To report our results we would provide the reader with descriptive statistics including the mean and standard deviation for each of the groups. Then we would report the degrees of freedom of the numerator and denominator of the F ratio, as well as the value of the F ratio that was obtained. If the F ratio was statistically significant, we would also provide a measure of effect size for the treatment. Since there are only two groups there would be no need to conduct post hoc comparisons in order to identify which groups differ. Thus, with our example we would report, "The sample means were found to differ significantly ($F(1,8) = 9.00$, $p < .05$, $\eta^2 = .53$)." As you will see, by using SPSS we can obtain greater accuracy and we can provide the p-value.

Progress Check

1. If there is a statistically significant treatment effect, the mean square between (MS_{Bet}) estimate of σ_X^2 will be _____ than the mean square within (MS_W) estimate, and the F ratio will be greater than _____.
2. The numerator of the F ratio is the _____ while the denominator of the F ratio is the _____.
3. A measure of effect size for a one-way between-subjects ANOVA is _____.

Answers: 1. Greater; 1 2. (MS_{Bet}) ; (MS_W) 3. eta squared

A Second Example

Let us assume that a group of researchers wants to study what effect information about the benefits of exercise has on weight loss in dieters. The researchers choose a design that consists of a control group and two experimental groups. The control group does not receive any special intervention. Subjects in the first experimental group receive written materials highlighting the benefits of engaging in an exercise program. The second experimental group is treated similarly to the first experimental group except that for the second experimental group there is an informational meeting instead of the written materials. The null hypothesis is that the populations from which these three samples are drawn do not differ in weight loss and thus have equal means. We will set the experimentwise α at .05. For each participant the number of pounds lost during the following year, the dependent measure, is recorded in Table 11.8, along with the deviation from the appropriate sample mean, and the squared deviation from the sample mean. Note that the sample sizes are unrealistically small and that they do not have to be equal.

Table 11.8 Example 2: Pounds Lost, Initial Calculations

Control Group (X_1)			Experimental Group I (X_2)			Experimental Group II (X_3)		
X_1	$(X_1 - M_1)$	$(X_1 - M_1)^2$	X_2	$(X_2 - M_2)$	$(X_2 - M_2)^2$	X_3	$(X_3 - M_3)$	$(X_3 - M_3)^2$
7	–5.60	31.36	10	–5.20	27.04	19	–3.00	9.00
11	–1.60	2.56	14	–1.20	1.44	20	–2.00	4.00
11	–1.60	2.56	15	–0.20	0.04	24	2.00	4.00

16	3.40	11.56	18	2.80	7.84	25	3.00	9.00
18	5.40	29.16	19	3.80	14.44			
$\Sigma X_1 = 63$	$\Sigma x_1 = 0$	$\Sigma x_1^2 = 77.20$	$\Sigma X_2 = 76$	$\Sigma x_2 = 0$	$\Sigma x_2^2 = 50.80$	$\Sigma X_3 = 88$	$\Sigma x_3 = 0$	$\Sigma x_3^2 = 26.00$
n = 5			n = 5			n = 4		
$M_1 = \frac{\mathbf{63}}{\mathbf{5}}$			$M_2 = \frac{\mathbf{76}}{\mathbf{5}}$			$M_3 = \frac{\mathbf{88}}{\mathbf{4}}$		
=12.60			= 15.20			= 22.00		

The next step is to create an ANOVA summary table showing what must be calculated (Table 11.4).

As each value is determined it is entered into Table 11.11. We will begin by finding the values for SS.

Calculating The Sums Of Squares

To find the SS we start with the equation:

$$SS_T = \Sigma(X - M_G)^2$$

where M_G, the grand mean, is the mean of all of the scores. It is found using the following equation:

$$M_G = \frac{\Sigma \mathbf{X}}{\mathbf{N}}$$

For our example with a total of 14 scores the calculation of M_G and SS_T is shown in Table 11.9.

Table 11.9 Example 2: Calculation of the Sums of Squares Total

X	$(X - M_G)$	$(X - M_G)^2$
7	−9.21	84.82
11	−5.21	27.14
11	−5.21	27.14
16	−0.21	0.04
18	1.79	3.20
10	−6.21	38.56
14	−2.21	4.88
15	−1.21	1.46
18	1.79	3.20
19	2.79	7.78
19	2.79	7.78
20	3.79	14.36
24	7.79	60.68
25	8.79	77.26

$\Sigma X = 227$ $\Sigma(X - M_G) = 0$ $\Sigma(X - M_G)^2 = 358.30 = SS_T$

$N = 14$

$$M_G = \frac{227}{14}$$

$$= 16.21$$

The SS_{Bet} is found by, first, determining the square of the deviation of a sample mean from the grand mean. This is then multiplied by the sample size. Finally, these steps are repeated for each sample mean and the resulting values are summed. Conceptually:

$$SS_{Bet} = \Sigma\,[(M - M_G)^2\, n]$$

where in this case $M_G = 16.21$, and n is the size of each sample. For our example the calculations are shown in Table 11.10.

Table 11.10 Example 2: Calculation of SS_{Bet}

	M	$(M - M_G)$	$(M - M_G)^2$	$(M - M_G)^2 n$
	Becomes:	$(M - 16.21)$	$(M - 16.21)^2$	$(M - 16.21)^2(n)$
Control	12.60	–3.61	13.03	(13.03)(5) = 65.15
Exp 1	15.20	–1.01	1.02	(1.02)(5) = 5.10
Exp 2	22.00	5.79	33.52	(33.52)(4) = 134.08
				$\Sigma(M - M_G)^2\, n = 204.33 = SS_{Bet}$

Alternatively, the same calculations can be presented as follows:

$$SS_{Bet} = \Sigma\,[(M - M_G)^2\, n]$$
$$= [(12.60 - 16.21)^2\,(5)] + [(15.20 - 16.21)^2\,(5)] + [(22.00 - 16.21)^2\,(4)]$$
$$= [(-3.61)^2\,(5)] + [(-1.01)^2\,(5)] + [(5.79)^2\,(4)]$$
$$= (13.03)(5) + (1.02)(5) + (33.52)(4)$$
$$= 65.15 + 5.10 + 134.08$$
$$= 204.33$$

The SS_W can be found by subtracting each score from its sample mean and squaring this deviation. The sum of these squared deviations for all of the scores in each of the groups would be SS_W . Thus, conceptually:

$$SS_W = \Sigma[\Sigma\,(X - M)^2]$$

where M is the mean of a group.

In other words, with three groups:

$$SS_W = \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2$$

Fortunately, we have already calculated these values in Table 11.8. Therefore:

$$SS_W = 77.20 + 50.80 + 26.00$$
$$= 154.00$$

In order to check our calculations we make use of the fact that:

$$SS_T = SS_{Bet} + SS_W$$
$$358.30 \approx 204.33 + 154.00$$
$$358.30 \approx 358.33$$

Note that the slight discrepancy is due to minor rounding error in our calculations.

Calculating The Degrees Of Freedom

We now must calculate the degrees of freedom for between groups, within groups and total, and enter these values in Table 11.11. The degrees of freedom for between groups is equal to the number of groups minus 1. Thus:

$$df_{Bet} = k - 1$$

where k is the number of levels of the IV.

In our example:

$$df_{Bet} = 3 - 1$$
$$= 2$$

To find the degrees of freedom for within groups we can first subtract 1 from the total number of subjects in a group and then sum across all of the groups. Thus:

$$df_W = \Sigma(n - 1)$$

where n is the number of subjects in each group or sample.

In our example:

$$df_W = (5 - 1) + (5 - 1) + (4 - 1)$$
$$= 4 + 4 + 3$$
$$= 11$$

Alternatively, we could find df_W by subtracting the number of groups from the total number of subjects:

$$df_W = N - k$$
$$= 14 - 3$$
$$= 11$$

To find the degrees of freedom total, we subtract 1 from the total number of subjects. Thus:

$df_T = N - 1$

where N is the total number of subjects in all the groups or samples.

In our example:

$df_T = 14 - 1$

$= 13$

As a check on our calculations:

$df_T = df_{Bet} + df_W$

$13 = 2 + 11$

$13 = 13$

Calculating The Mean Squares

The MS between groups and the MS within groups are found by dividing the appropriate SS by its degrees of freedom. Thus:

$MS_{Bet} = \frac{\mathbf{SS_{Bet}}}{\mathbf{df_{Bet}}}$ $\quad$ $MS_W = \frac{\mathbf{SS_W}}{\mathbf{df_W}}$

$= \frac{\mathbf{204.33}}{\mathbf{2}}$ $\quad$ $= \frac{\mathbf{154.00}}{\mathbf{11}}$

$= 102.17$ $\quad$ $= 14.00$

Calculating The F Ratio

Finally, we calculate our F ratio:

$F = \frac{\mathbf{MS_{Bet}}}{\mathbf{MS_W}}$

$= \frac{\mathbf{102.17}}{\mathbf{14.00}}$

$= 7.30$

The ANOVA summary table is now complete (Table 11.11). We will later calculate the value listed in the final column of this table.

Table 11.11 Example 2: Completed Summary Table for the One-way Between-subjects ANOVA, with the Value for Eta Squared (η^2)

Source of Variation	SS	df	MS	F	η^2
Between Groups	204.33	2	102.17	7.30	.57
Within Groups	154.00	11	14.00		
Total	358.30	13			

Interpreting The F Ratio

To determine whether this F ratio of 7.30 is significantly different from a value of 1, which you will recall is what would be expected if the independent variable had no effect, we must enter the F table (Appendix K, Table 4). Remember, the F ratio is based upon two MS estimates of the population variance, each with its degrees of freedom. To find the critical value of F we locate the column corresponding to the degrees of freedom associated with the numerator of the F ratio and the row corresponding to the degrees of freedom associated with the denominator of the F ratio. For our F this would be 2 and 11 degrees of freedom. The critical value of F with the α of .05 is 3.98. As our obtained F of 7.30 is larger than the critical value, we reject the null hypothesis that the samples came from populations with equal means and accept the alternative hypothesis that the population means differ. Put differently, we conclude that our independent variable had an effect on our dependent variable.

Conducting The Post Hoc Comparisons

While a significant F indicates that the independent variable had an effect, with three or more levels of the IV it does not specify which group means differ. It was noted previously that the number of pairwise comparisons in an experiment is given by the equation:

Number of pairwise comparisons $= \frac{\mathbf{k(k-1)}}{\mathbf{2}}$

where k is the number of groups, samples or treatment levels

In our case, as k equals 3 there are [3(3 – 1)] / 2, which equals 3, pairwise comparisons. Specifically, the 3 pairwise comparisons are between the mean of the Control Group and the mean of Experimental Group I, the mean of the Control Group and the mean of Experimental Group II, and the mean of Experimental Group I and the mean of Experimental Group II. The significant F indicates that *at least one* of the group means is expected to differ from another. To specify which means differ we need to conduct what are called post hoc tests. You are familiar with the concept of post hoc tests for we used them following a significant chi-square test in Chapter 8 when the chi-square design was larger than a 2 X 2.

A researcher can choose from a number of post hoc tests that are used after a significant F ratio is found. One of the most popular and easiest to calculate is Tukey's honestly significant difference (**Tukey HSD**) test.

<u>Tukey HSD</u> – A popular post hoc test used with ANOVAs.

Calculation of the Tukey HSD leads to a critical value that is compared to the difference of each of the post hoc pairwise comparisons of group means in the study. Specifically:

Critical value of Tukey HSD $= \text{q} \sqrt{\frac{\mathbf{MS_W}}{\mathbf{n}}}$

where q is found from the q Table (Appendix K, Table 5). The column to use is determined by the number of means being compared (the number of levels of the IV), in our case 3. The row is determined by the degrees of freedom of the MS_W, in our case 11. With α equal to .05, q is equal to 3.82. *(Be careful, this is not the critical value for the Tukey HSD test.)*

The value of MS_W comes from Table 11.11. It is equal to 14.00.

The value of n equals the number of subjects in *each* group if the number of subjects in each group is the same.

Alternatively:

$$n = \frac{\textbf{Number of means}}{\sum \frac{1}{\textbf{Number of subjects in each group}}}$$

if the group size is not the same for all of the groups.

In our example we do not have an equal number of subjects in each sample. We therefore calculate the n for use in finding the critical value as follows:

$$\begin{aligned} n &= \frac{\mathbf{3}}{\frac{\mathbf{1}}{\mathbf{5}}+\frac{\mathbf{1}}{\mathbf{5}}+\frac{\mathbf{1}}{\mathbf{4}}} \\ &= \frac{\mathbf{3}}{\mathbf{0.20 + 0.20 + 0.25}} \\ &= \frac{\mathbf{3}}{\mathbf{0.65}} \\ &= 4.62 \end{aligned}$$

To find the critical value, we now substitute into the equation:

$$\text{Critical value of Tukey HSD} = q \sqrt{\frac{\mathbf{MS_W}}{\mathbf{n}}}$$

We found that MS_W equals 14.00 and n has just been calculated. The value for q is found in the q table. We can now find the critical value for the Tukey HSD:

$$\begin{aligned} \text{Critical value} &= 3.82 \sqrt{\frac{\mathbf{14.00}}{\mathbf{4.62}}} \\ &= 3.82 \sqrt{3.03} \\ &= (3.82)\ (1.74) \\ &= 6.65 \end{aligned}$$

The difference between the means for each pairwise comparison must be *as great or greater* than the critical value from the Tukey HSD in order to be considered statistically significant:

Difference between the means of the Control Group and Experimental Group I

$= 12.60 – 15.20 = –2.60$

Difference between the means of the Control Group and Experimental Group II

$= 12.60 – 22.00 = –9.40$

Difference between the means of Experimental Group I and Experimental Group II

$= 15.20 – 22.00 = –6.80$

It is important to note that when comparing the differences between group means (in our case –2.60, –9.40 and –6.80) to the critical value, we ignore the sign as this simply reflects the order in which the sample means were subtracted. Our critical value is 6.65. Therefore, the difference between the means of the Control Group and Experimental Group I, which is 2.60, is not significant. However, the difference between the means of the Control Group and Experimental Group II, which is 9.40, and the difference between the means of Experimental Group I and Experimental Group II, which is 6.80, are both statistically significant.

Calculating The Effect Size

We now proceed to ascertain the percent of variance explained by the treatment. To do so we calculate eta squared (η^2). With a one-way between-subjects ANOVA:

$$\eta^2 \text{ for treatment} = \frac{\mathbf{SS_{Bet}}}{\mathbf{SS_T}}$$

This is a measure of the proportion of total variability explained or accounted for by the treatment. In our example:

$$\eta^2 \text{ for treatment} = \frac{\mathbf{204.33}}{\mathbf{358.30}} = .57 \text{ or } 57\%$$

This value of η^2 is included in the last column of Table 11.11.

In addition, we can calculate η^2 for the error term in the ANOVA. This is the proportion of variability *not* accounted for by the treatment:

$$\eta^2 \text{ for error} = \frac{\mathbf{SS_W}}{\mathbf{SS_T}}$$

$$= \frac{\mathbf{154.00}}{\mathbf{358.30}}$$

$$= .43 \text{ or } 43\%$$

The η^2 values for treatment and error will sum to 1.00. In our case .57 + .43 = 1.00.

Reporting The Results Of A One-Way Between-Subjects ANOVA

In a paper, we would provide the mean and standard deviation for each of the groups. Then we would report the degrees of freedom of the numerator and denominator of the F ratio, as well as the value of the F ratio that was obtained. If the F ratio was statistically significant, we would provide a measure of effect size for the treatment, and we would identify which pairwise comparisons were statistically significant. Specifically, based upon our calculations we would report, "The sample means were found to differ significantly ($F(2,11) = 7.30$, $p < .05$, $\eta^2 = .57$)." We would then indicate that Tukey's HSD test indicated that the control group lost less weight than the second experimental group, but not the first, and the second experimental group lost more weight than the first experimental group.

With SPSS we can give a precise p-value. For the overall ANOVA we would report (F(2,11) = 7.30, p = .010, η^2 = .57). Note that the p-value of .010 is less than our α of .05, confirming that we would reject the null hypothesis.

Purpose And Limitations Of Using The One-way Between-subjects ANOVA

1. *Test for difference.* The null hypothesis is that the treatment does not have an effect. Therefore, if the null is correct any difference between the group means is due to chance. The alternative hypothesis is that the treatment does have an effect and, therefore, the samples are drawn from populations with different means. The one-way between-subjects ANOVA is employed to differentiate between these two hypotheses.
2. *Does not provide a measure of effect size.* The one-way between-subjects ANOVA is a test of significance. It indicates whether or not an outcome is likely to have occurred by chance if the null hypothesis is correct. If the F test is significant, a measure of effect size, such as eta squared (η^2), should then be calculated.
3. *Compares two or more group means.* The one-way between-subjects ANOVA is appropriate to use when the independent variable has two or more levels and when each subject is randomly assigned to only one level of the independent variable.
4. *Does not indicate where the effect is.* With designs with more than two levels to the independent variable, a significant F should be followed by a post hoc procedure such as the Tukey HSD test in order to specify the location of the effect.

Assumptions Of The One-way Between-subjects ANOVA

1. *Interval or ratio data.* The data are on an interval or ratio scale of measurement.
2. *Random samples.* Each sample is drawn at random from a population.
3. *Independence within treatment levels.* The data within each treatment level are independent.
4. *Normally distributed populations.* Each population from which a sample is drawn has a normal distribution of scores. However, as stated in the Central Limit Theorem, the F test will be accurate so long as each sample size is at least 30. If a sample size is less than 30 then it is important that the underlying population be normally distributed. If you cannot collect a larger sample and do not know if the assumption of normality has been met, it may be best to turn to an alternative test on the same row of Table 11.1 that does not assume that the data are normally distributed.
5. *Population variances are equal.* The populations from which samples are drawn have equal variances.

Conclusion

The logic of the one-way between-subjects ANOVA is based upon examining the study's variability from two perspectives. One perspective is based upon finding the variability of scores within each group and then summing across all the groups. This leads to an estimate of the population variability called MS_W. Since the subjects within each group receive the same level of the independent variable this estimate of the population variability does not include any effect of treatment. It only reflects the variability in the study that is not due to treatment, which is called error.

The other perspective is to examine the variability of the sample means. This estimate is called MS_{Bet}. If there is no treatment effect we still expect the sample means to vary, somewhat if the underlying population does not have a great deal of variability, more if the underlying population does have a great deal of variability. (Sample size would also have to be taken into account.) However MS_{Bet}, unlike MS_W, may be affected by treatment. Specifically, if the IV has an effect then MS_{Bet}, but not MS_W, will increase.

It can be shown that if there is no treatment effect then these two estimates of the population variance will be approximately equal and thus the F ratio will be approximately equal to 1.00. However, if there is a treatment effect then MS_{Bet} will be greater than MS_W and the F ratio will be greater than 1.00.

If the F ratio is found to be statistically significant then a measure of effect size such as eta squared should be calculated. And, if the independent variable has more than two levels then the Tukey HSD test should be utilized to determine where the effect is.

Future chapters will review two additional forms of the ANOVA. All of these procedures are closely related. However, it is important to keep them distinct. Appendix M provides a summary of the similarities and differences of these three ANOVAs.

Final Thoughts: The Relationship Between The t Test And The F Test

The one-way between-subjects ANOVA is a very flexible test and serves as an introduction to the more complex ANOVAs that will be covered in subsequent chapters. The major advantage of the ANOVA compared to the independent samples t test is that the ANOVA controls the experimentwise error rate while simultaneously comparing two or more sample means.

Though the calculations for the independent samples t test (reviewed in Chapter 10) and the one-way between-subjects ANOVA with two groups appear to be quite different, these tests are actually closely related. In fact, the independent samples t test is a special case of the one-way between-subjects ANOVA. It is possible, for instance, when there are only two groups to convert the outcome obtained with one test into the value of the other using the following equation:

$F = t^2$

Previously, we found a value of F = 9.00 (Table 11.7). Using the equation above, we have:

$F = 9.00 = t^2$

Therefore, t = 3.00. You are encouraged to verify that this is indeed the case by redoing the example as a t test.

Since the values obtained with the t test and F test for two groups can be converted into each other, it should not be a surprise that these tests will always lead to the same decision as to whether the null hypothesis should be retained or rejected.

Similarly, the estimate of the effect size will be the same regardless of whether you calculate an independent samples t test or the one-way between-subjects ANOVA with two groups. You are encouraged to verify that this is the case for our examples where the IV had two levels.

Finally, though it is probably not immediately obvious, the t and F tables are also closely related. Unlike the t table which requires only knowing the degrees of freedom derived from the number of subjects, the F table requires that you know two degrees of freedom because the F ratio is based on two MS estimates, MS_{Bet} and MS_W, each with its degrees of freedom. The degrees of freedom for MS_W, like the degrees of freedom for the t table, reflects the number of subjects. The degrees of freedom for the MS_{Bet} is based on the number of means being compared. With the independent samples t test the number of means being compared is always two. With the one-way between-subjects ANOVA it is two or more. This affects the critical value and thus must be accounted for in the F table. However, if we are comparing only two means (the df_{Bet} in the ANOVA would then equal k – 1 = 1) then the values in the F table equal the square of the values in the t table for a two-tailed test (remember $F = t^2$). Thus, with $\alpha = .05$, the critical value for F with $df_{Bet} = 1$ and $df_W = 1$ is 161.45. For t with a two-tailed test and 1 df the critical value is 12.71. We therefore have:

$F = t^2$

$161.45 = 12.71^2$

161.45 = 161.54 except for minor rounding error

And with $\alpha = .05$, the critical value for F with $df_{Bet} = 1$ and $df_W = 5$ is 6.61 and the critical value for t for a two-tailed test with 5 df is 2.57. We therefore have:

$F = t^2$

$6.61 = 2.57^2$

6.61 = 6.60 except for minor rounding error

You are encouraged to verify that this relationship holds for other degrees of freedom for the t test and the matching df_W for the ANOVA. Just remember that the df_{Bet} for the ANOVA must remain equal to 1.

Glossary of Terms

Analysis of variance (ANOVA) – A set of flexible, closely related, inferential procedures for comparing sample means by examining variances.

Between-subjects design – With an ANOVA, those designs in which each subject experiences only a single level of a factor.

Error – With ANOVA, the variability not due to treatment.

Experimentwise error rate – The likelihood of making at least one Type I error with any of the experiment's comparisons.

Factor – With an ANOVA, the term 'Factor' is often used instead of independent variable.

Factorial ANOVA – An ANOVA with more than one factor.

Grand mean (M_G) – The mean of the sample means. In some statistical procedures it is defined as the mean of all of the scores.

Level – With an ANOVA, the number of values of an independent variable.

Mean square (MS) – In an ANOVA, an estimate of the population variance (σ_X^2).

Mean square between (MS_{Bet}) – The estimate of the population variance (σ_X^2) based upon the variability between the sample means. More specifically, it is obtained from the deviations of the sample means from the grand mean.

Mean square within (MS_W) – The estimate of the population variance (σ_X^2) based upon the variability within each of the samples. More specifically, it is obtained by pooling the variances of the scores within each of the samples.

One-way between-subjects ANOVA – An inferential procedure for comparing two or more means from independent samples when there is one independent variable.

Pairwise comparison – Comparison between two sample means.

Pairwise error rate – The likelihood of making a Type I error for a single comparison between sample means. This is equal to α, which is usually .05 or .01.

Sum of squares between groups (SS_{Bet}) – The sum of the squared deviations of each treatment mean from the grand mean.

Sum of squares total (SS_T) – The sum of the squared deviations from the mean for all of the scores.

Sum of squares within groups (SS_W) – The sum across all conditions, of the sum of the squared deviations of each score from its treatment mean.

<u>Treatment</u> – With ANOVA, another term for the independent variable.

<u>Tukey HSD</u> – A popular post hoc test used with ANOVAs.

References

Brown, M. B. & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association, 69,* 364-367. (Chap 11, in SPSS)

Questions – Chapter 11

(Answers are provided in Appendix J.)

1. The sum of squares total in a one-way between-subjects ANOVA is equal to _____.
   a. The sum of all of the squared scores
   b. The sum of all the scores, squared
   c. The sum of the squared deviations of all of the scores from the grand mean
   d. The square root of the sum of the squared deviations of all of the scores from the grand mean

2. In a one-way between-subjects ANOVA, SS_T is equal to _____.
   a. SS_W
   b. SS_{Bet}
   c. $SS_{Bet} + SS_W$
   d. $SS_{Bet} – SS_W$

3. 'Within variability' provides an estimate of _____ by looking at _____.
   a. Control group variability; how much sample means vary
   b. Population variability; how much scores vary from their sample means
   c. Experimental group(s) variability; how much sample means vary
   d. How much sample means vary; population variability

4. Another term for 'mean square' is _____.
   a. Variance
   b. Standard deviation
   c. Range
   d. Square of the sum of all of the sample means

5. If there is no treatment effect, the F ratio is expected to approximately equal _____.
   a. 0
   b. 1
   c. 2
   d. Twice the number of experimental conditions

6. If the F ratio from a one-way between-subjects ANOVA with two levels is found to be statistically significant, the researcher should _____.

a. Announce that the finding is important
b. Conduct a post-hoc test to determine which groups differ
c. Calculate eta squared
d. None of the above, there is nothing further to do

7. The F ratio can only be significant if it is _____.
a. Less than 1
b. Equal to 1
c. Greater than 1
d. An F ratio can never be significant

8. If the F ratio from a one-way between-subjects ANOVA with more than two levels is found to be statistically significant, the researcher should _____.
a. Announce that the finding is important
b. Conduct a post-hoc test to determine which groups differ
c. Re-calculate as it is obvious that an error has occurred
d. None of the above, there is nothing further to do

9. In a one-way between-subjects ANOVA there are how many independent variables and how many dependent variables?
a. 1; 1
b. 1; 2
c. 2; 1
d. 2; 2

Questions 10 to 14 are based upon the following table:

Source of Variation	SS	df	MS	F
Between Groups	____	2	___	___
Within Groups	100	___	___	
Total	400	12		

10. What is the SS for Between Groups?
a. 300
b. 500
c. 40,000
d. 100

11. What is the value of df for Within Groups?
a. 14
b. 10
c. 33
d. 26

12. What is MS for Between Groups?
a. 150
b. 600
c. 298
d. 502

13. What is MS for Within Groups?
a. 1000
b. 50
c. 25

d. 10

14. What is the value of F?
a. 60
b. 15
c. 2.5
d. 1

Questions 15 to 19 are based upon the following information:

A group of 18 students take their final exam in statistics. Each student is randomly assigned to one of three rooms; quiet, moderately noisy or noisy. The number of errors for each student is:

<u>Level of Background Noise</u>

Quiet	Moderate	Noisy
9	7	6
10	9	8
8	8	10
13	13	7
12	11	11
14	12	12

15. What is the df_{Bet}?
a. 1
b. 2
c. 3
d. 4

16. What is the SS_{W}?
a. 12
b. 36
c. 84
d. 96

17. What is the MS_{Bet}?
a. 6
b. 12
c. 36
d. 146

18. What is the value of F?
a. 0
b. 26.07
c. 0.92
d. 1.07

19. Is the outcome statistically significant with alpha equal to .05?
a. yes
b. no

Questions 20 and 21 deal with the relationship of F and t.

20. It is possible to convert the outcome obtained with the t test into the value of a Between Subjects ANOVA using the equation _____.
  a. F = 2t
  b. $t^2 = F$
  c. t/6 = 10F
  d. t = F

21. The t and F tests for studies with two groups will _____ lead to the same decision as to whether the null hypothesis should be retained or rejected.
  a. Always
  b. Sometimes
  c. Never

Problems 22 – 29 utilize SPSS.

Using SPSS With The One-Way Between-Subjects ANOVA

To Begin SPSS

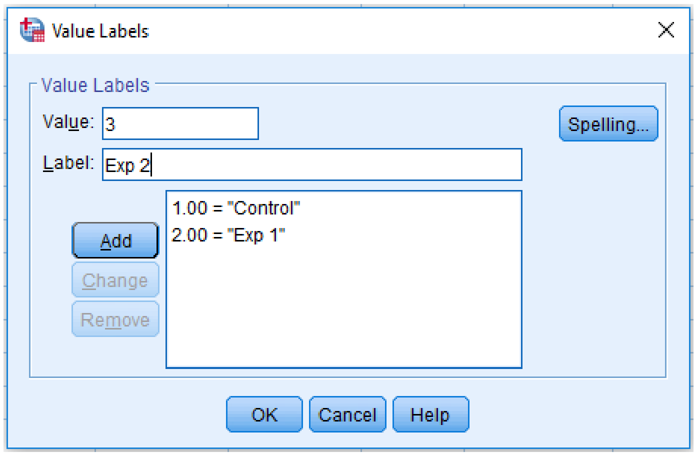
Step 1 Activate the program, close the central window, and click on the **Variable View** option at the bottom left of the window.

Step 2 Click on the first empty rectangle (called a 'cell') under the column heading 'Name'. You now type the name of the first variable for which you have data. We are going to utilize the same data and labels as were previously employed in Table 11.8. These data dealt with the question of whether receiving information about the benefits of exercise would affect weight loss. We have called these variables 'Condition' and 'Data'. Therefore, type 'Condition' in the first empty cell under 'Name'.

Step 3 Click on the first empty 'cell' under the column heading 'Label', and type 'Experimental Group'.

Step 4 Click on the first empty 'cell' under the column heading 'Values'. A box will appear. In the blank space to the right of 'Value', type the number '1'. Then type a brief description of this value of the variable in the blank space to the right of 'Label'. In our case, type 'Control'. Finally, click on 'Add'. Your label for a value of 1 will appear in the large white region in the center of the window. Now repeat the above steps in this section for the value '2', which is given the label 'Exp 1', and for the value '3', which is given the label 'Exp 2' (Figure 11.1). Click 'Add' and then click on 'OK'.

Figure 11.1 The Value Labels Window



Value: 3

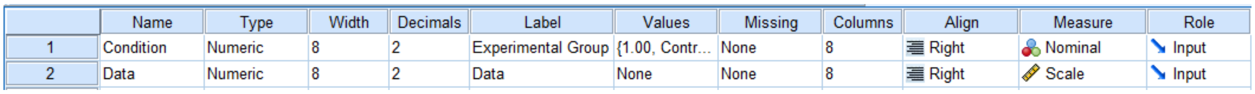
Label: Exp 2

1.00 = "Control"
2.00 = "Exp 1"

Step 5 Click on the first empty 'cell' under the column heading 'Measure'. As we are dealing with labels for groups, select 'Nominal'.

Step 6 Repeat Steps 2 – 5 except that you type 'Data' in the first empty cell under 'Name' and for the 'Label'. Finally, select 'Scale' in the column under the column heading 'Measure' as we have ratio data. The result is shown in Figure 11.2.

Figure 11.2 The Variable View Window



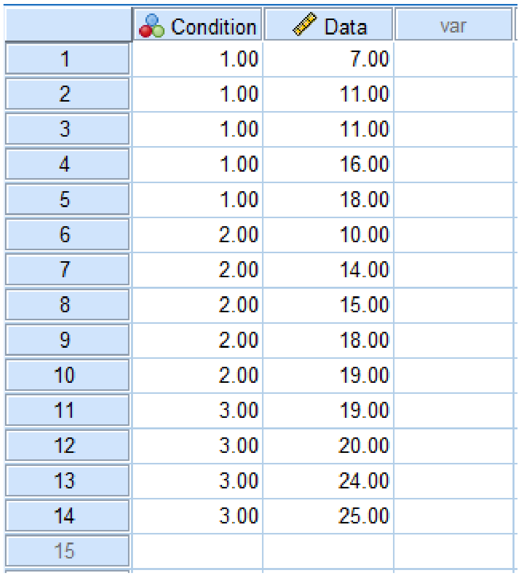
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Condition	Numeric	8	2	Experimental Group	{1.00, Contr...	None	8	Right	Nominal	Input
2	Data	Numeric	8	2	Data	None	None	8	Right	Scale	Input

To Enter Data In SPSS

Step 7 Click on the 'Data View' option at the lower left corner of the window. The variables 'Condition' and 'Data' will be present.

Step 8 For each subject in the control condition, type the value '1' in the column 'Condition' and their weight loss in the column 'Data' (Figure 11.4). Continue by entering '2' for each subject in group 2 with their data and finally '3' for each subject in group 3 with their data (Figure 11.3).

Figure 11.3 Entering the Data



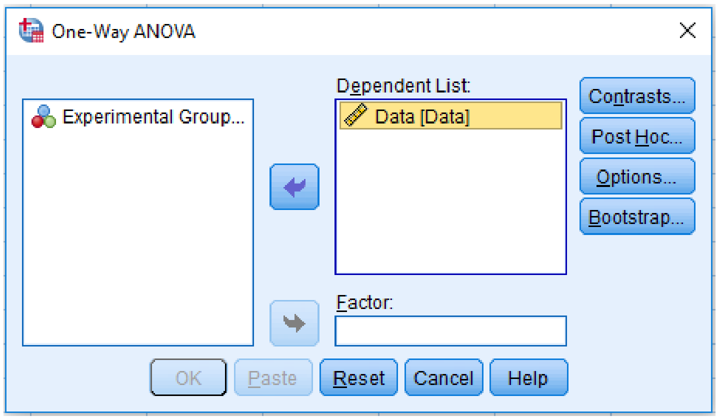
	Condition	Data	var
1	1.00	7.00	
2	1.00	11.00	
3	1.00	11.00	
4	1.00	16.00	
5	1.00	18.00	
6	2.00	10.00	
7	2.00	14.00	
8	2.00	15.00	
9	2.00	18.00	
10	2.00	19.00	
11	3.00	19.00	
12	3.00	20.00	
13	3.00	24.00	
14	3.00	25.00	
15			

To Conduct A One-way Between-subjects ANOVA

Step 9 Click the cursor on ‘**Analyze**’ along the row of SPSS commands above the data you entered, then move to ‘**Compare Means**’, then click on ‘**One-way ANOVA**’.

Step 10 A new window will appear. This asks for the dependent variable and the independent variable (called a Factor) to be identified. In our case, Data is the label of the dependent variable. This is indicated by moving ‘Data’ to the box under ‘Dependent list’ by clicking on ‘Data’ and then on the top arrow. The result is shown in Figure 11.4. Then move ‘Experimental Group’ to the box under ‘Factor’ by clicking on ‘Experimental Group’ and then on the bottom arrow. The result will be that each label will now be in the appropriate box on the right-hand side of the window, as is shown in Figure 11.5. Then click on ‘**Post Hoc**’ which is located in the column on the right of the window.

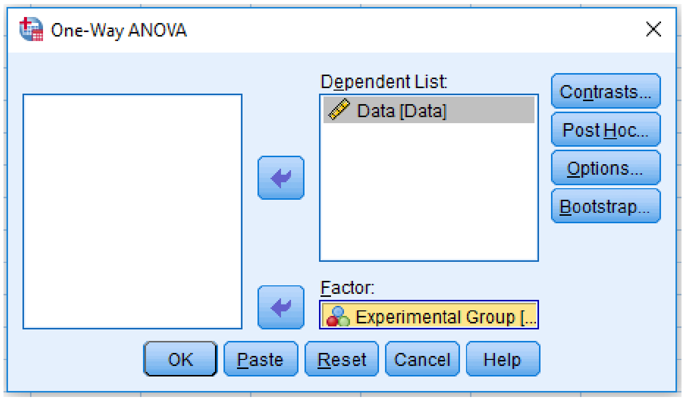
Figure 11.4 The One-way ANOVA Window



Dependent List:
Data [Data]

Factor:

Figure 11.5 The One-way ANOVA Window, Continued



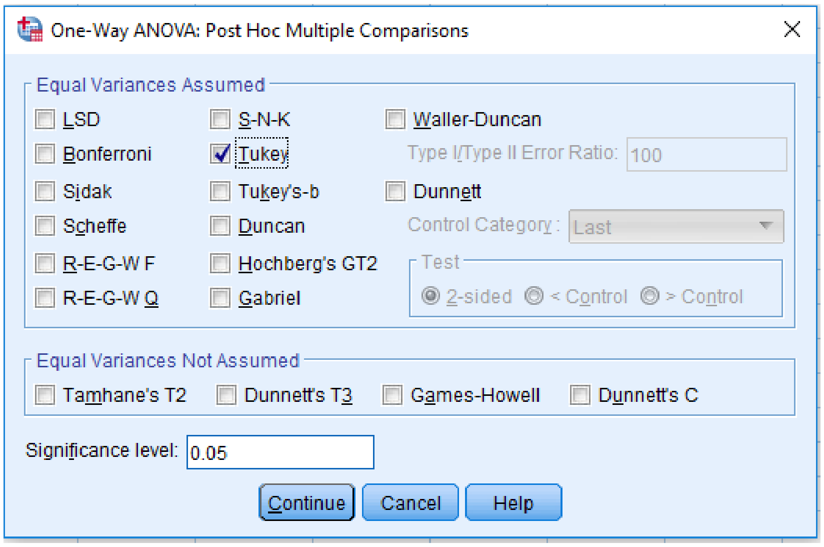
One-Way ANOVA

Dependent List: Data [Data]

Factor: Experimental Group [...

Step 11 A new window will appear. This window provides a number of statistical options that are available with SPSS. In this book we will limit ourselves to just the Tukey HSD test. Click on '**Tukey**' as it is shown in Figure 11.6. Then click on '**Continue**'.

Figure 11.6 Identifying the Post Hoc Test



One-Way ANOVA: Post Hoc Multiple Comparisons

Equal Variances Assumed

☐ LSD ☐ S-N-K ☐ Waller-Duncan

☐ Bonferroni ☑ Tukey Type I/Type II Error Ratio: 100

☐ Sidak ☐ Tukey's-b ☐ Dunnett

☐ Scheffe ☐ Duncan Control Category: Last

☐ R-E-G-W F ☐ Hochberg's GT2 Test

☐ R-E-G-W Q ☐ Gabriel ● 2-sided ○ < Control ○ > Control

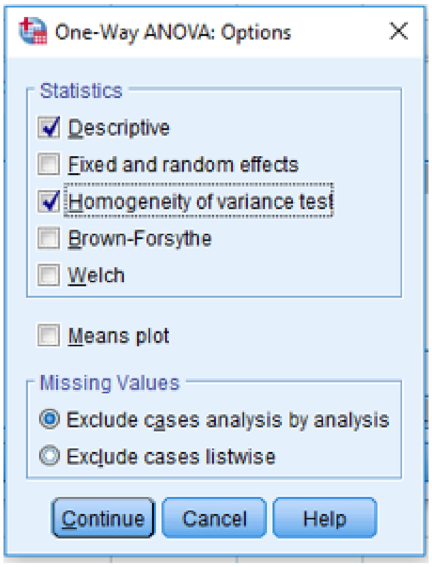
Equal Variances Not Assumed

☐ Tamhane's T2 ☐ Dunnett's T3 ☐ Games-Howell ☐ Dunnett's C

Significance level: 0.05

Step 12 Now click on '**Options**' which is located in the column on the right of the window. A new window will appear. If you click on the boxes in front of '**Descriptive**' and '**Homogeneity of variance test**' (Figure 11.7), SPSS will later generate a useful summary of the data and calculate Levene's test for homogeneity of variances. Click on '**Continue**'.

Figure 11.7 Specifying Descriptives



One-Way ANOVA: Options

Statistics

- ☑ Descriptive
- □ Fixed and random effects
- ☑ Homogeneity of variance test
- □ Brown-Forsythe
- □ Welch

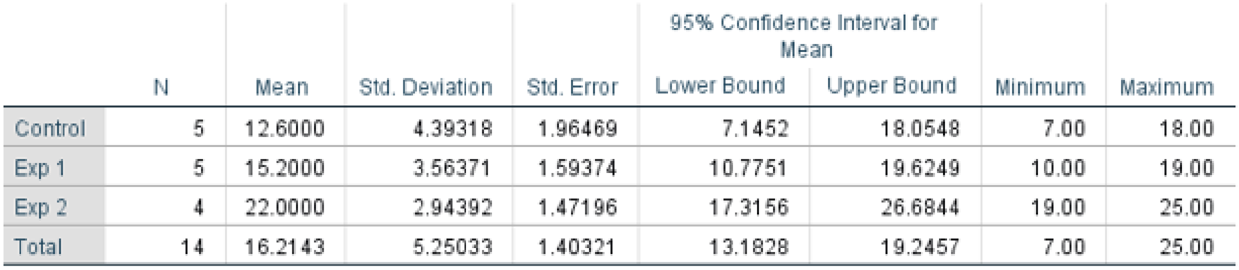
□ Means plot

Missing Values

- ● Exclude cases analysis by analysis
- ○ Exclude cases listwise

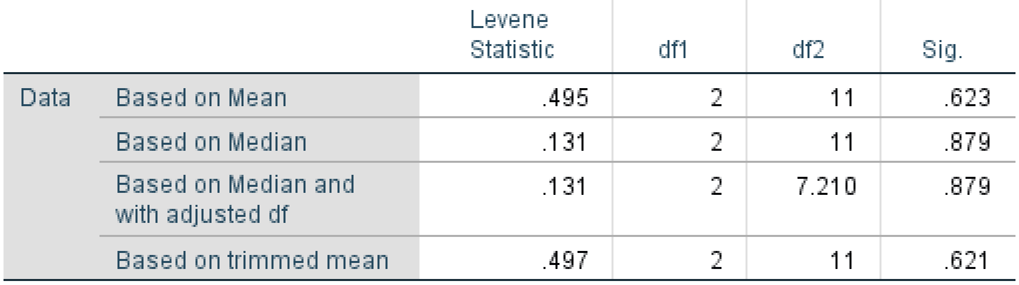
Step 13 Now click on '**OK**'. SPSS calculates the desired one-way ANOVA with descriptive statistics, the test for homogeneity of variance, and the Tukey HSD post hoc test as is shown in Tables 11.12 – 11.16.

Table 11.12 SPSS Output; One-way ANOVA – Descriptives and Confidence Intervals



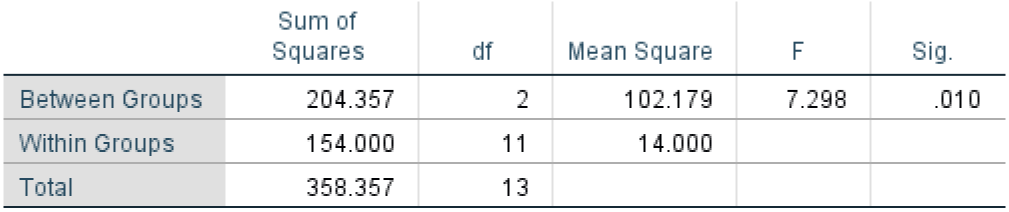
					95% Confidence Interval for Mean			
	N	Mean	Std. Deviation	Std. Error	Lower Bound	Upper Bound	Minimum	Maximum
Control	5	12.6000	4.39318	1.96469	7.1452	18.0548	7.00	18.00
Exp 1	5	15.2000	3.56371	1.59374	10.7751	19.6249	10.00	19.00
Exp 2	4	22.0000	2.94392	1.47196	17.3156	26.6844	19.00	25.00
Total	14	16.2143	5.25033	1.40321	13.1828	19.2457	7.00	25.00

Table 11.13 SPSS Output; Test of Homogeneity of Variances



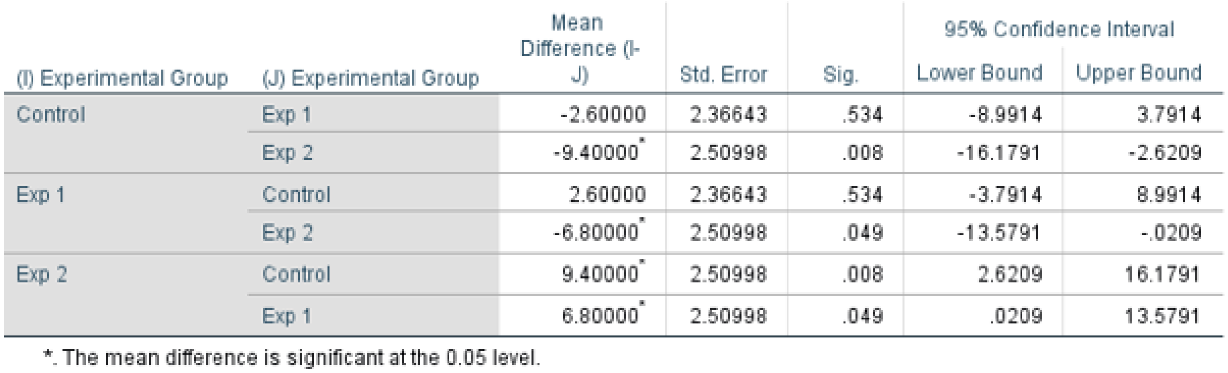
		Levene Statistic	df1	df2	Sig.
Data	Based on Mean	.495	2	11	.623
	Based on Median	.131	2	11	.879
	Based on Median and with adjusted df	.131	2	7.210	.879
	Based on trimmed mean	.497	2	11	.621

Table 11.14 SPSS Output; One-way ANOVA Summary Table



	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	204.357	2	102.179	7.298	.010
Within Groups	154.000	11	14.000		
Total	358.357	13			

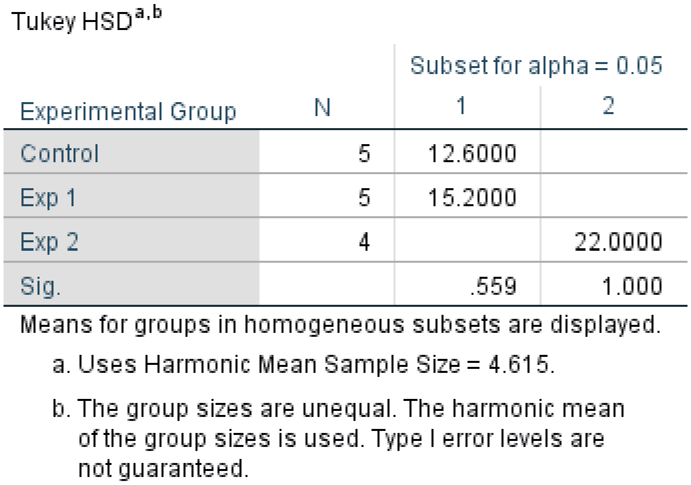
Table 11.15 SPSS Output; Tukey HSD Multiple Comparisons



(I) Experimental Group	(J) Experimental Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Control	Exp 1	-2.60000	2.36643	.534	-8.9914	3.7914
	Exp 2	-9.40000*	2.50998	.008	-16.1791	-2.6209
Exp 1	Control	2.60000	2.36643	.534	-3.7914	8.9914
	Exp 2	-6.80000*	2.50998	.049	-13.5791	-.0209
Exp 2	Control	9.40000*	2.50998	.008	2.6209	16.1791
	Exp 1	6.80000*	2.50998	.049	.0209	13.5791

*. The mean difference is significant at the 0.05 level.

Table 11.16 SPSS Output; Alternative Presentation of Tukey HSD Multiple Comparisons



Tukey HSD[a,b]

Experimental Group	N	Subset for alpha = 0.05	
		1	2
Control	5	12.6000	
Exp 1	5	15.2000	
Exp 2	4		22.0000
Sig.		.559	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 4.615.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

The first section of the analysis is a table of descriptive statistics (Table 11.12). You should compare the means of the samples that we calculated by hand (Table 11.8) with the means calculated by SPSS. In addition, standard deviations, 95% confidence intervals, and additional information that might be of interest are included. Next is Levene's test for homogeneity of variances (Table 11.13). A number of options are provided. In general, Levene's test based upon the median is recommended (Brown & Forsythe, 1974). As the significance (p-value) is .879, which is greater than .05, we conclude that the variances of the populations from which the three groups were selected are not significantly different and thus we can continue with the ANOVA. If the obtained significance had been less than .05 then we should turn to an alternative statistical procedure, such as the Kruskal Wallis H test which is reviewed in Appendix A. The summary ANOVA table, which is the same as we found earlier with hand calculation (Table 11.11) except that we previously rounded off our calculations to fewer places, is shown next (Table 11.14). The fourth table shows the results of the Tukey HSD post hoc test (Table 11.15). In the first two rows of the analysis, the Control Group is compared with the two experimental groups and the difference with Experimental Group 2 is found to be statistically significant (This is indicated with a small * in the Mean Difference column) as well as by a significance level (p-value) less than .05 (in this case .008). In addition, 95% confidence intervals are provided. These confidence intervals are for the difference between group means. Any interval that does not include 0 will indicate a statistically significant difference. The next two pairs of rows provide the comparisons for Experimental Group 1 and Experimental Group 2. The results are the same as we found previously by hand. Table 11.16 provides an alternative presentation of the Tukey HSD output. Specifically, the sample sizes are given, and which sample means differ are clearly indicated by the column in which they are listed. Thus, listing the Control and Exp 1 groups in the same column indicates they do not differ. However, Exp 2 is in a separate column which indicates it differs from both the Control and Exp 1 groups. We do not need to concern ourselves with the last row of this table.

While SPSS provides a great deal of information it does not provide a value for eta squared (η^2). Fortunately, η^2 is easy to calculate from the information in Table 11.14:

$$\eta^2 \text{ for treatment} = \frac{\mathbf{SS_{Bet}}}{\mathbf{SS_T}} = \frac{\mathbf{204.357}}{\mathbf{358.357}} = .570 \text{ or } 57\%$$

This is the same value we obtained with our previous hand calculations (Table 11.11).

Step 14 Exit SPSS. There is no need to save the output or data.

To confirm that you understand how to use SPSS, I suggest you redo the between-subjects ANOVA that was calculated in the text for the data in Table 11.3, but this time using SPSS. Then redo the ANOVA dealing with level of background noise (Questions 15 – 19) to check your answers.

SPSS Problems – Chapter 11

For questions 22 – 29, what is the effect of adding a constant (in this case 10) to every score in the noisy condition of the data used for questions 15 – 19? (Compare your answers for these data to your previous answers.)

	Level of Background Noise	
Quiet	Moderate	Noisy
9	7	16
10	9	18

8	8	20
13	13	17
12	11	21
14	12	22

22. Is the significance of Levene’s test of homogeneity of variance less than .05?
    a. yes
    b. no

23. What is the df_{Bet}?
    a. 1
    b. 2
    c. 3
    d. 4

24. What is the SS_W?
    a. 12
    b. 36
    c. 84
    d. 96

25. What is the MS_{Bet}?
    a. 6
    b. 12
    c. 36
    d. 146

26. What is the value of F?
    a. 0
    b. 26.07
    c. 0.92
    d. 1.07

27. Is the outcome statistically significant with alpha equal to .05?
    a. yes
    b. no

28 If you had found the significance of Levene’s test was .02 you would _____.
    a. Continue to conduct the ANOVA
    b. Use the independent samples t test instead
    c. Turn to the Kruskal-Wallis H test
    d. Stop and not do any further analysis

29. If you had found the significance of Levene’s test was .42 you would_____.
    a. Continue to conduct the ANOVA
    b. Use the independent samples t test instead
    c. Turn to the Kruskal-Wallis H test
    d. Stop and not do any further analysis

Chapter 12
Finding Differences with Interval/Ratio Data – IV: The One-way Within-subjects ANOVA

"…the null hypothesis is never proved or established, but is possibly disproved,
in the course of experimentation."
R. A. Fisher

Logic of a One-way Within-subjects ANOVA

The logic of the **one-way within-subjects ANOVA** builds upon what we already learned for the one-way *between-subjects* ANOVA. Both include one independent variable (IV) with two or more levels, one dependent variable (DV), and both culminate in the calculation of an F ratio. To review, F is the ratio of two estimates of the population variance, σ_X^2. With the one-way *between-subjects* ANOVA, the estimate in the *numerator* is based on the *variability of the sample means*. This estimate of σ_X^2 is called the mean square between (MS_{Bet}). It includes the effect of our treatment as well as sample variation unrelated to any treatment effect, which is called error.

One-way within-subjects ANOVA – An inferential procedure for comparing two or more means from related samples when there is one independent variable.

And with the one-way *between-subjects* ANOVA the estimate of the population variance (σ_X^2) in the *denominator* of the F ratio is obtained by noting how much each *score varies from its sample mean*. Thus this estimate does not include the effect of our treatment since each subject in a sample receives the same level of the treatment. The combined estimate of σ_X^2 from looking within each of the samples in the study is called the mean square within (MS_W). It is a measure of the variation in the data excluding any treatment effect (this is called error).

For the one-way *between-subjects* ANOVA the F ratio is:

$$F = \frac{\mathbf{MS_{Bet}}}{\mathbf{MS_W}} = \frac{\textbf{Estimate of } \sigma_X^2 \textbf{ based upon treatment + error}}{\textbf{Estimate of } \sigma_X^2 \textbf{ based only upon error}}$$

If there is no treatment effect, then F will become the ratio of two estimates of the population variance based solely upon what we are calling error. These two estimates should be approximately equal, and thus the ratio would be about 1.00. If, on the other hand, there is a treatment effect then the numerator will be greater than the denominator and the F ratio will be greater than 1.00.

Thus far we have discussed two sources of variability, treatment and error. However, you will see shortly that there are actually two types of error. Therefore, the outcome of a study can potentially be impacted by a total of three sources of variability. One source of variability is the level of the treatment that the subject receives. If the treatment has an effect, then the behavior of subjects will differ depending upon which treatment level they were assigned to. This is what the experimenter is interested in determining. Unfortunately, there are two other sources of variability, collectively known as error, which can make this determination difficult. One source of error reflects differences in relatively stable characteristics of subjects, such as their heights or IQ. These are called **preexisting subject differences** since they are characteristics of the subjects before the study even begins. The other source of error is due to events (e.g., changes in the temperature, whether a subject became ill, or if a subject was just accepted into graduate school) that happen to coincide with the testing. This variability is called **residual error**.

The presence of error (both preexisting subject differences and residual error) can make it difficult for the experimenter to determine whether their treatment had an effect. As you will see, the one-way *within-subjects* ANOVA is a popular statistical procedure because it eliminates the pre-existing subject differences from the analysis. This removal of one of the components of the error in a study can often increase the likelihood that the experimenter will be able to ascertain whether their treatment had an effect. The one-way within-subjects ANOVA is underlined in Table 12.1.

Preexisting subject differences – Relatively stable subject characteristics. These differences between subjects are a form of error in an ANOVA. The variability due to these differences is removed in a one-way within-subjects ANOVA.

Residual error – Changeable subject characteristics. These differences between subjects are a form of error in an ANOVA. The variability due to these differences is not removed in a one-way within-subjects ANOVA.

Table 12.1 Overview of Inferential Statistical Procedures For Finding if there is a Difference

	Type of Data			
	Nominal (Frequency)		Ordinal (Ranked)	Interval/Ratio (Continuous Measure)
Research Design		Research Design		
One Variable With At Least Two Outcomes	Goodness-of-fit Chi-Square	One IV With One Sample		One-sample z Test or One-sample t Test

		One IV With Two Or More Independent Samples	*Kruskal–Wallis H*	One-way Between–Subjects ANOVA (Only two independent samples, Independent Samples t Test)
		One IV With One Sample Having Two Or More Repeated Measures		One-way Within–Subjects ANOVA (Only two repeated measures, Dependent Samples t Test)
Two Variables, Each With At Least Two Outcomes	Chi-Square Test of Independence	Two IV Each With Two Or More Independent Samples		Two-way Between–Subjects ANOVA

The Italicized procedure is reviewed in Appendix A

Let us assume that we randomly select three samples of the same size from a population that has a great deal of variability. The variability within each of the samples would presumably be quite large, as would be the MS_W. This is because MS_W is an estimate of the population variance based upon the variability within each of the samples. It reflects preexisting subject differences and residual error, but not treatment.

How could we reduce the part of error that is due to preexisting subject differences? One solution would be to use repeated measures of the same subjects. With this procedure, since the same subjects are tested at each treatment level the same preexisting subject differences will also occur at each treatment level in the study. Stated differently, none of the variability *between* the treatment levels could then be due to preexisting subject differences. Thus, as you will see, with a repeated measures design the preexisting subject differences can be eliminated as a source of error. In other words, as the same subjects are now being tested at each treatment level, the treatment level means should be identical except for the effects of the treatment and residual error. The following discussion will make this point clear.

We previously noted that for the one-way *between-subjects* ANOVA, the F ratio is:

$$\mathrm{F} = \frac{\mathbf{MS_{Bet}}}{\mathbf{MS_W}} = \frac{\textbf{Estimate of } \sigma^2_X \textbf{ based upon treatment + error}}{\textbf{Estimate of } \sigma^2_X \textbf{ based only upon error}}$$

We have now found that there are two types of error which are called preexisting subject differences and residual error. Therefore, the F ratio can be rewritten as:

$$\mathrm{F} = \frac{\textbf{Estimate of } \sigma^2_X \textbf{ based upon treatment + preexisting subject differences + residual error}}{\textbf{Estimate of } \sigma^2_X \textbf{ based on preexisting subject differences + residual error}}$$

With the one-way *within-subjects* ANOVA (also known as the single-factor within-subjects ANOVA, the single-factor repeated measures ANOVA or the one-way repeated measures ANOVA), the preexisting subject differences are eliminated as a source of error from both the numerator and denominator of the F ratio. As a result, the F ratio becomes:

$$\mathrm{F} = \frac{\textbf{Estimate of } \sigma_X^2 \textbf{ based upon treatment} + \cancel{\textbf{preexisting subject differences}} + \textbf{residual error}}{\textbf{Estimate of } \sigma_X^2 \textbf{ based on } \cancel{\textbf{preexisting subject differences}} + \textbf{residual error}}$$

$$= \frac{\textbf{Estimate of } \sigma_X^2 \textbf{ based upon treatment} + \textbf{residual error}}{\textbf{Estimate of } \sigma_X^2 \textbf{ based only on residual error}}$$

Eliminating preexisting subject differences can have a dramatic effect upon the F ratio. For instance, using the equation based upon the one-way *between-subjects* ANOVA, if the estimate of treatment variance is 20, the preexisting subject differences estimate is 10, and the residual error estimate is 5, the F ratio is:

$$\mathrm{F} = \frac{\textbf{Estimate of } \sigma_X^2 \textbf{ based upon treatment} + \textbf{preexisting subject differences} + \textbf{residual error}}{\textbf{Estimate of } \sigma_X^2 \textbf{ based on preexisting subject differences} + \textbf{residual error}}$$

$$= \frac{\mathbf{20 + 10 + 5}}{\mathbf{10 + 5}}$$

$$= \frac{\mathbf{35}}{\mathbf{15}}$$

$$= 2.33$$

With a one-way *within-subjects* ANOVA the variability due to the preexisting subject differences is eliminated from the analysis and the F ratio becomes:

$$\mathrm{F} = \frac{\textbf{Estimate of } \sigma_X^2 \textbf{ based upon treatment} + \cancel{\textbf{preexisting subject differences}} + \textbf{residual error}}{\textbf{Estimate of } \sigma_X^2 \textbf{ based on } \cancel{\textbf{preexisting subject differences}} + \textbf{residual error}}$$

$$\mathrm{F} = \frac{\textbf{Estimate of } \sigma_X^2 \textbf{ based upon treatment} + \textbf{residual error}}{\textbf{Estimate of } \sigma_X^2 \textbf{ based only on residual error}}$$

$$= \frac{\mathbf{20} + \cancel{\mathbf{10}} + \mathbf{5}}{\cancel{\mathbf{10}} + \mathbf{5}}$$

$$= \frac{\mathbf{20 + 5}}{\mathbf{5}}$$

$$= 5.00$$

Clearly, use of the one-way *within-subjects* ANOVA design can dramatically increase the magnitude of the F ratio and, therefore, assist a researcher in detecting whether an independent variable has had an effect.

Before we turn to our first example it is important to note when this elimination of preexisting subject differences occurs. As was just noted, the preexisting subject differences are being removed from both the numerator and denominator of the F ratio with a repeated measures design. The numerator of the F ratio reflects differences in the dependent variable (DV) *between* the treatment levels. However, as was previously noted, with a repeated measures design these differences cannot be due to stable (preexisting) subject differences since the same subjects are being tested at each treatment level. In other words, the variability due to preexisting subject differences is eliminated from the numerator of the F ratio as a result of using the same subjects repeatedly. Thus, the elimination of the preexisting subject differences from the numerator of the F ratio is a consequence of the assignment of the same subjects to each of the treatment levels. This occurs at an early stage of the study, before calculation of the ANOVA. In contrast, the elimination of the preexisting subject differences from the denominator of the F ratio, which reflects differences in the DV *within* the treatment levels, is accomplished later, during the actual calculation of the ANOVA. This distinction is further explained in our first example.

Conducting A One-Way Within-Subjects ANOVA

The one-way *within-subjects* ANOVA can be used to analyze interval or ratio data from designs with two or more repeated measures. For our first example we will examine whether a fuel additive changes car mileage (this problem was also analyzed with the dependent samples t test in Chapter 10). The null hypothesis is that the fuel additive does not affect mileage. The alternative hypothesis is that it does. The alpha level is set at .05.

It is critical when using ANOVAs that you know what you are calculating and that you keep your calculations clearly defined. It is thus important that we begin with a table showing what it is that must be calculated (Table 12.2). You will recognize that the table for the one-way within-subjects ANOVA is similar, but not identical, to the table that we used with the one-way between-subjects ANOVA (Table 11.4).

Table 12.2 Summary Table for the One-way Within-subjects ANOVA

Source of Variation	SS	df	MS	F
Between Treatments	SS_{Bet}	df_{Bet}	MS_{Bet}	F ratio
Subjects	$SS_{Subjects}$	$df_{subjects}$		
Residual	$SS_{Residual}$	$df_{Residual}$	$MS_{Residual}$	
Total	SS_T	df_T		

A value must be recorded for each of the eleven entries underlined in Table 12.2. Thus, we will begin by finding four values for SS, and then calculate four values for df, two values for MS and one F ratio. You will see that this involves a substantial amount of calculation but, just as with the one-way between-subjects ANOVA, the advantage of the current ANOVA is that, unlike the dependent samples t test, it can be used with experimental designs that have more than two levels of the independent variable. And no step is difficult. It is critical, however, that you clearly identify each item you are calculating and enter the result in Table 12.7 so you do not become confused.

The data, which consist of two measurements of mileage for each car, the steps leading to the calculation of the SS, as well as the treatment means are reproduced in Table 12.3. (We will be referring to these calculations shortly.)

Table 12.3 Example 1: Initial Calculations

'Subjects'	Vehicle Mileage with	Mileage without	Subject
Cars	Additive (1)	Additive (2)	Totals

	X_1	$(X_1 – M_1)$	$(X_1 – M_1)^2$	X_2	$(X_2 – M_2)$	$(X_2 – M_2)^2$	$\Sigma X_{Subject}$
1	13	–4.50	20.25	12	–5.00	25.00	25
2	15	–2.50	6.25	13	–4.00	16.00	28
3	14	–3.50	12.25	15	–2.00	4.00	29
4	17	–0.50	0.25	17	0.00	0.00	34
5	24	6.50	42.25	20	3.00	9.00	44
6	22	4.50	20.25	25	8.00	64.00	47
	$\Sigma X_1 = 105$	$\Sigma x_1 = 0$	$\Sigma x_1^2 = 101.5$	$\Sigma X_2 = 102$	$\Sigma x_2 = 0$	$\Sigma x_2^2 = 118$	$\Sigma\ (\Sigma X_{Subject}) = 207$
	$n = 6$			$n = 6$			
	$M_1 = 105 / 6$			$M_2 = 102 / 6$			
	$= 17.50$			$= 17.00$			

Calculating The Sums Of Squares

Our first step, as with the one-way between-subjects ANOVA, is to find each of the sums of squares (SS). Specifically, it was noted in Chapter 11 that:

$SS_T = SS_{Bet} + SS_W$

In the one-way within-subjects ANOVA, the sum of squares within (SS_W) is partitioned into the **sum of squares subjects** ($SS_{Subjects}$), which is the SS due to preexisting subject differences, and the **sum of squares residual** ($SS_{Residual}$), which is the SS due to residual error. Thus, in a one-way within-subjects ANOVA:

$SS_T = SS_{Bet} + SS_{Subjects} + SS_{Residual}$

Sum of squares subjects ($SS_{Subjects}$) – In a one-way within-subjects ANOVA, the SS due to preexisting subject differences.

Sum of squares residual ($SS_{Residual}$) – In a one-way within-subjects ANOVA, the SS due to residual error.

We begin our calculations by finding SS_T in the same way as with a one-way between-subjects ANOVA:

$SS_T = \Sigma(X – M_G)^2$

where M_G is the mean of all of the scores.

The grand mean (M_G) is found using the following equation:

$M_G = \frac{\mathbf{\Sigma X}}{\mathbf{N}}$

where N is the total number of data points or scores.

For our example with a total of 12 scores from 6 subjects (i.e., cars), calculation of the grand mean is shown in the first column of Table 12.4. Calculation of the SS_T is shown in the remaining columns of Table 12.4.

Table 12.4 Example 1: Calculation of M_G and SS_T

X	$(X – M_G)$	$(X – M_G)^2$
13	–4.25	18.06
15	–2.25	5.06
14	–3.25	10.56
17	– 0.25	0.06
24	6.75	45.56
22	4.75	22.56
12	–5.25	27.56
13	–4.25	18.06
15	–2.25	5.06
17	– 0.25	0.06
20	2.75	7.56
25	7.75	60.06
$\Sigma X = 207$	$\Sigma(X – M_G) = 0$	$\Sigma(X – M_G)^2 = 220.22 = SS_T$
$N = 12$		
$M_G = \frac{\mathbf{207}}{\mathbf{12}}$		
$= 17.25$		

The SS_{Bet} for a one-way within-subjects ANOVA is found, as in a one-way between-subjects ANOVA, by determining the square of the deviations of each treatment level mean from the grand mean, multiplying by the number of subjects in the sample, and then summing. It is important to note that while the equation and thus the computations used to determine the SS_{Bet} are the same for the between-subjects and within-subjects ANOVAs, this term is, nevertheless, interpreted somewhat differently depending upon which ANOVA is being utilized. The SS_{Bet} for a *between-subjects ANOVA* is referring to variability between the *groups* in the study. As was noted previously, this variability reflects that each group receives a different treatment level and that each group is composed of different subjects. It thus includes preexisting subject differences as well as residual error. However, in a *repeated measures* (*within-subjects) design* the treatment means cannot be affected by differences in the composition of the groups since the same subjects (in this case, cars) receive each of the treatment levels. Thus, the same preexisting subject differences exist within each of the treatment levels. In other words there is no variability *between treatment levels* due to preexisting subject differences. Consequently, *the SS_{Bet} for a one-way within-subjects ANOVA does not include variability due to preexisting subject differences. It only reflects the effect of the treatment as well as any residual error.* The equation for SS_{Bet} remains:

$$SS_{Bet} = \Sigma[(M - M_G)^2 n]$$

Recall that the means of our two samples are 17.50 and 17.00. The grand mean (M_G) equals 17.25 and n, the number of subjects (cars), equals 6. For our example the calculations to determine SS_{Bet} are shown in Table 12.5.

Table 12.5 Example 1: Calculation of SS_{Bet}

	M	$(M - M_G)$	$(M - M_G)^2$	$(M - M_G)^2n$
	Becomes:	$(M - 17.25)$	$(M - 17.25)^2$	$(M - 17.25)^2(6)$
w Additive	17.50	0.25	0.06	0.36
wo Additive	17.00	−0.25	0.06	0.36
				$\Sigma(M - M_G)^2 n = 0.72 = SS_{Bet}$

Alternatively, the same calculations can be presented as follows:

$$SS_{Bet} = \Sigma[(M - M_G)^2 n]$$
$$= [(17.50 - 17.25)^2(6)] + [(17.00 - 17.25)^2(6)]$$
$$= [(0.25)^2(6)] + [(-0.25)^2(6)]$$
$$= (0.06)(6) + (0.06)(6)$$
$$= 0.36 + 0.36$$
$$= 0.72$$

As with a one-way between-subjects ANOVA, the value of SS_W can be found using the following equation:

$$SS_W = \Sigma x_1^2 + \Sigma x_2^2$$

These values of SS have already been calculated in Table 12.3:

$$SS_W = 101.50 + 118.00$$
$$= 219.50$$

And we use the following equation to check our calculations:

$$SS_T = SS_{Bet} + SS_W$$
$$220.22 = 0.72 + 219.50$$

220.22 = 220.22

The value of SS_W is <u>not</u> entered into our ANOVA table for in a within-subjects ANOVA it must be partitioned into $SS_{Subjects}$ and $SS_{Residual}$.

The $SS_{Subjects}$, which is the part of SS_W that is due to preexisting subject differences, is found by determining the deviation of the mean for a *subject* from the grand mean, squaring this deviation, multiplying by the number of treatment levels, and then summing for each subject. Conceptually:

$$SS_{Subjects} = [\Sigma(\frac{\mathbf{\Sigma X_{Subject}}}{\mathbf{k}} - \mathbf{M_G})^2\mathbf{k}]$$
$$= [\Sigma(\mathbf{M_{Subject}} - \mathbf{M_G})^2\mathbf{k}]$$

where the subject's total ($\Sigma X_{Subject}$) is obtained from Table 12.3, the grand mean (M_G) is obtained from Table 12.4 and k is the number of treatment levels (in this example, k = 2). The calculation of $SS_{Subjects}$ is shown in Table 12.6.

Table 12.6 Example 1: Calculation of $SS_{Subjects}$

'Subjects' Cars	$\Sigma X_{Subject}$	$\mathbf{M_{Subject}} = \frac{\mathbf{\Sigma X_{Subject}}}{\mathbf{k}}$	$\mathbf{M_{Subject}} - \mathbf{M_G}$	$(\mathbf{M_{Subject}} - \mathbf{M_G})^2$	$(\mathbf{M_{Subject}} - \mathbf{M_G})^2\mathbf{k} = (\mathbf{M_{Subject}} - \mathbf{M_G})^2(\mathbf{2})$
1	25	12.50	–4.75	22.56	45.12
2	28	14.00	–3.25	10.56	21.12
3	29	14.50	–2.75	7.56	15.12
4	34	17.00	–0.25	0.06	0.12
5	44	22.00	4.75	22.56	45.12
6	47	23.50	<u>6.25</u>	39.06	<u>78.12</u>
			$\Sigma(\mathbf{M_{Subject}} - \mathbf{M_G}) = \mathbf{0}$		$\Sigma(\mathbf{M_{Subject}} - \mathbf{M_G})^2(\mathbf{2})$ = 204.72 = $SS_{Subjects}$

Alternatively, the same calculations can be presented as follows:

$$SS_{Subjects} = [\Sigma(\mathbf{M_{Subject}} - \mathbf{M_G})^2\mathbf{k}]$$
$$= [(12.50 – 17.25)^2](2) + [(14.00 – 17.25)^2](2) + [(14.50 – 17.25)^2](2) + [(17.00 – 17.25)^2](2) + [(22.00 – 17.25)^2](2) + [(23.50 – 17.25)^2](2)$$
$$= [(–4.75)^2](2) + [(–3.25)^2](2) + [(–2.75)^2](2) + [(–0.25)^2](2) + [(4.75)^2](2) + [(6.25)^2](2)$$
$$= (22.56)(2) + (10.56)(2) + (7.56)(2) + (0.06)(2) + (22.56)(2) + (39.06)(2)$$
$$= 45.12 + 21.12 + 15.12 + 0.12 + 45.12 + 78.12$$

$= 204.72$

Remember that SS_W is being partitioned into $SS_{Subjects}$ and $SS_{Residual}$. The value for $SS_{Subjects}$ is entered into our ANOVA summary table. The $SS_{Subjects}$ is the reduction to the denominator of the F ratio achieved by using a within-subjects design.

As was noted previously, in a within-subjects ANOVA,

$$SS_W = SS_{Subjects} + SS_{Residual}$$

Substituting:

$$219.50 = 204.72 + SS_{Residual}$$

$$14.78 = SS_{Residual}$$

This value for $SS_{Residual}$ is entered into the within-subjects ANOVA summary table. And you will soon see that $SS_{Residual}$ is used in calculating the denominator of the F ratio.

Calculating Degrees Of Freedom

We now must calculate our degrees of freedom. Fortunately, these values are all easy to obtain.

The degrees of freedom for between treatments is equal to the number of treatment levels minus 1. Thus:

$$df_{Bet} = k - 1$$

where k is the number of treatment levels. In our example:

$$df_{Bet} = 2 - 1$$

$$= 1$$

In a one-way *between-subjects* ANOVA we would now determine the df_W and enter the value in the summary table. In a one-way *within-subjects* ANOVA, df_W is partitioned into $df_{Subjects}$ and $df_{Residual}$, just as SS_W was partitioned into $SS_{Subjects}$ and $SS_{Residual}$. The degrees of freedom for subjects is equal to the *number of subjects* minus 1. (Note that the data consist of 12 mileages, but there are only 6 subjects, in this case cars.) Thus:

$$df_{Subjects} = n - 1$$

where n is the number of subjects. In our example:

$$df_{Subjects} = 6 - 1$$

$$= 5$$

As you will see, the $df_{Subjects}$ is the number of degrees of freedom that are removed from the analysis due to using a within-subjects design.

To find the degrees of freedom for residual, we subtract 1 from the total number of subjects and multiply this by the number of levels minus 1. Thus:

$$df_{Residual} = (n - 1)(k - 1)$$

where n is the number of subjects and k is the number of levels. In our example:

$$\begin{aligned} df_{Residual} &= (6 - 1)(2 - 1) \\ &= (5)(1) \\ &= 5 \end{aligned}$$

To find the degrees of freedom for total, we subtract 1 from the total number of data points. Thus:

$$df_T = N - 1$$

where N is the total number of data points. In our example:

$$\begin{aligned} df_T &= 12 - 1 \\ &= 11 \end{aligned}$$

As a check on our calculations:

$$df_T = df_{Bet} + df_{Subjects} + df_{Residual}$$

$$11 = 1 + 5 + 5$$

$$11 = 11$$

These values for df are entered in our ANOVA summary table.

Calculating Mean Squares

You will recall that in a one-way *between-subjects* ANOVA (reviewed in Chapter 11) the value of the F ratio is obtained by dividing the estimate of the population variance derived from variability of the group means (MS_{Bet}) by the estimate of the population variance derived from variability of scores within each group (MS_W). In a one-way *within-subjects* ANOVA we also calculate an F ratio. The calculation of the MS_{Bet} is the same for both types of ANOVA. However, as noted previously, the numerator of a within-subjects ANOVA does not include the variability due to preexisting subject differences since the same subjects are being tested at each level of the IV. And, in a one-way *within-subjects* ANOVA the $SS_{Subjects}$ has been partitioned out of the SS_W leaving only $SS_{Residual}$ in the denominator of the F ratio. Thus, both the numerator and the denominator are being reduced by eliminating the variability due to preexisting subject differences.

As in a one-way between-subjects ANOVA, the F ratio for a one-way within-subjects ANOVA is the ratio of two estimates of population variability. However, in the *within-subjects* ANOVA the F ratio is determined by dividing MS_{Bet} by $MS_{Residual}$, not MS_W as was the case for the *between-subjects* ANOVA. And remember that each mean square is found by dividing the appropriate SS by its df. Thus:

$$MS_{Bet} = \frac{\mathbf{SS_{Bet}}}{\mathbf{df_{Bet}}} = \frac{\mathbf{0.72}}{\mathbf{1}} = 0.72 \qquad MS_{Residual} = \frac{\mathbf{SS_{Residual}}}{\mathbf{df_{Residual}}} = \frac{\mathbf{14.78}}{\mathbf{5}} = 2.96$$

Calculating The F Ratio

The final calculation is to determine the value of the F ratio. As was just noted, the equation for the F ratio for a one-way *within-subjects* ANOVA is:

$$F = \frac{\mathbf{MS_{Bet}}}{\mathbf{MS_{Residual}}} = \frac{\mathbf{0.72}}{\mathbf{2.96}} = 0.24$$

Inclusion of the values for MS_{Bet}, $MS_{Residual}$, and F complete Table 12.7.

Table 12.7 Example 1: Completed Summary Table for the One-way Within-subjects ANOVA

Source of Variation	SS	df	MS	F
Between Treatments	0.72	1	0.72	0.24
Subjects	204.72	5		
Residual	14.78	5	2.96	
Total	220.22	11		

Interpreting The F Ratio

Recall that if the independent variable did not have an effect we would expect the F ratio to equal 1.00. And if there was a treatment effect then the value of the F ratio would be greater than 1.00. As our F ratio is less than 1.00 we know even without entering the F table that this outcome is not statistically significant. Nevertheless, if you wanted to use the F table to find the critical value of F you would locate the column corresponding to the degrees of freedom of the numerator of our F ratio and the row corresponding to the degrees of freedom of our denominator. From Table 12.7 we see that for our F this would be 1 and 5 degrees of freedom. We chose an α of .05. At the intersection of our column and row in the F table (Appendix K, Table 4) we find the critical value of

6.61. As our obtained F of 0.24 is less than the critical value we retain the null hypothesis that the fuel additive did not affect the gas mileage.

Since the F was not significant we do not calculate a measure of effect size. And even if the F was significant we would not conduct post hoc comparisons as this study had only two treatment levels and, consequently, we would already know the treatment levels that differed from each other.

Reporting The Results Of A One-Way Within-Subjects ANOVA Without A Significant F Ratio

In a paper, we would indicate the degrees of freedom as well as the F ratio that was obtained. Specifically, we would, based upon our calculations, report that the fuel additive was not found to affect vehicle mileage ($F(1,5) = 0.24$, $p > .05$). Note the direction of the > sign. Later in this chapter we will use SPSS to analyze these data, and we then can make a more precise statement ($F(1,5) = 0.25$, $p = .636$). The minor change in the value of the F ratio is due to rounding error in our calculations, and SPSS provides a precise p-value for the probability of our outcome. Note that the p-value of .636 is greater than our α of .05, confirming that we would retain the null hypothesis. (It is also instructive to note that the probability of our F ratio, which was .636, is the same as when these data were analyzed with SPSS using a dependent samples t test (Table 10.12). Different statistical procedures, but same probability and decision.)

Progress Check

1. Compared to a one-way *between-subjects* ANOVA, in a one-way *within-subjects* ANOVA the the variability due to preexisting subject differences is _____ from both the numerator and denominator of the F ratio.
2. The calculation of the one-way within-subjects ANOVA eliminates _____ subject differences from the denominator.
3. In a one-way within-subjects ANOVA, the denominator of the F ratio consists only of _____.

Answers: 1. eliminated 2. preexisting 3. residual error

A Second Example

Our next example will show that the one-way within-subjects ANOVA can be used when there are more than two measures from each subject. Let's assume that you are a researcher and you are interested in whether housing choice affects exam scores among students who excel. You decide to utilize a within-subjects design in which each student lives in three different housing situations; an on-campus honors dorm, off-campus at home, and off-campus in an apartment with other students, each for a 4-week period. As each student receives each treatment level, this is a repeated measures design. Your null hypothesis is that choice of housing will not affect exam scores, and you set α equal to .05. In order to control for order effects, which occur when a particular sequence of treatments has a unique outcome, you assign the sequence of these treatment levels randomly to each subject. At the end of each 4-week housing period you measure each subject's academic success with a 100-point exam. The hypothetical scores, squared deviations, treatment level means and each subject's total are presented in Table 12.8. Note that an unrealistically small sample size has been chosen to aid in the calculations.

Table 12.8 Example 2: Initial Calculations

Subject	On-campus			Off-campus			Off-campus			Subject
	Honors Dorm (1)			At Home (2)			Apartment (3)			Totals
	X_1	$(X_1 – M_1)$	$(X_1 – M_1)^2$	X_2	$(X_2 – M_2)$	$(X_2 – M_2)^2$	X_3	$(X_3 – M_3)$	$(X_3 – M_3)^2$	$\Sigma X_{Subject}$
1	98	4	16	96	3	9	84	2.8	7.84	278
2	96	2	4	95	2	4	81	–.2	0.04	272
3	95	1	1	95	2	4	82	.8	0.64	272
4	91	–3	9	91	– 2	4	80	–1.2	1.44	262
5	90	–4	16	88	–5	25	79	–2.2	4.84	257
	$\Sigma X_1 = 470$	$\Sigma x_1 = 0$	$\Sigma x_1^2 = 46$	$\Sigma X_2 = 465$	$\Sigma x_2 = 0$	$\Sigma x_2^2 = 46$	$\Sigma X_3 = 406$	$\Sigma x_3 = 0$	$\Sigma x_3^2 = 14.80$	$\Sigma\ (\Sigma X_{Subject}) = 1341$
	$n = 5$			$n = 5$			$n = 5$			
	$M_1 = 470\ /\ 5$			$M_2 = 465\ /\ 5$			$M_3 = 406\ /\ 5$			
	$= 94.00$			$= 93.00$			$= 81.20$			

Before proceeding, recall that with a one-way *between-subjects* ANOVA, we employed Levene's test when using SPSS to determine whether to maintain the assumption of homogeneity of variance of our groups. Levene's test is not utilized with a one-way *within-subjects* ANOVA. Instead, when we turn to SPSS we will be using **Mauchly's test** to examine a related assumption called **sphericity**. This assumption is that the variances of the differences between treatment levels are equal. In our current example, each subject lived in three different locations; an honors dorm, at home, and in an apartment. The differences between each subject's scores in the three housing situations can easily be calculated (Table 12.9). The assumption of sphericity is that the variances for these differences (columns D_1, D_2, and D_3) do not differ. When we turn to SPSS we will be using Mauchly's test to determine whether we should accept this assumption. For now, we will proceed as if we had conducted Mauchly's test and concluded that the assumption of sphericity would be retained. Finally, it is important to note that since sphericity is only of concern when there is more than one set of difference scores, it is only relevant when the IV has at least three levels.

Mauchly's test of sphericity – Statistical procedure utilized with SPSS to test the assumption of sphericity for a one-way within-subjects ANOVA.

Sphericity – Assumption of a within-subjects ANOVA that the variances of the sets of difference scores between treatment levels are equal. In a repeated measures ANOVA these differences would be based upon pairs of scores from each subject.

Table 12.9 Example 2: Differences in Students' Scores Based Upon Living Situation

Subject	Dorm	− Home	= D_1	Dorm	− Apartment	= D_2	Home	− Apartment	= D_3
1	98	96	2	98	84	14	96	84	12
2	96	95	1	96	81	15	95	81	14
3	95	95	0	95	82	13	95	82	13
4	91	91	0	91	80	11	91	80	11
5	90	88	2	90	79	11	88	79	9

We will be using the same summary table as in our previous example (Table 12.2) and will proceed by finding a value for each underlined item. As each value is calculated it is entered into Table 12.13..

Calculating The Sums Of Squares

We start by finding our SS. Remember:

$$SS_T = SS_{Bet} + SS_W$$

And with a one-way within-subjects ANOVA, $SS_W = SS_{Subjects} + SS_{Residual}$

As before, we determine the value of SS_T by using the following equation:

$$SS_T = \Sigma(X - M_G)^2$$

where M_G is the mean of all of the scores, in other words the grand mean. It is found using the following equation:

$$M_G = \frac{\Sigma \mathbf{X}}{\mathbf{N}}$$

where N is the total number of scores.

For our example with a total of 15 scores from 5 subjects the calculation of M_G is shown in the first column of Table 12.10. Calculation of SS_T is shown in the remaining columns of Table 12.10.

Table 12.10 Example 2: Calculation of M_G and SS_T

X	$(X - M_G)$	$(X - M_G)^2$
98	8.60	73.96

96	6.60	43.56
95	5.60	31.36
91	1.60	2.56
90	0.60	0 .36
96	6.60	43.56
95	5.60	31.36
95	5.60	31.36
91	1.60	2.56
88	–1.40	1.96
84	–5.40	29.16
81	–8.40	70.56
82	–7.40	54.76
80	–9.40	88.36
79	–10.40	108.16
$\Sigma X = 1341$	$\Sigma(X - M_G) = 0$	$\Sigma(X - M_G)^2 = 613.60 = SS_T$

$N = 15$

$M_G = 1341 / 15$

$= 89.40$

The SS_{Bet} is found as in a between-subjects ANOVA by determining the square of the deviations of each treatment level mean (M) from the grand mean (M_G), then multiplying by the number of subjects (n), and finally summing. Thus:

$$SS_{Bet} = \Sigma[(M - M_G)^2 n]$$

Recall, however, that with a within-subjects ANOVA this calculation does not include variability due to preexisting subject differences.

The treatment level means (M_1 through M_3) come from Table 12.8, M_G equals 89.40 (Table 12.10) and n, the number of subjects, is 5. The calculations for SS_{Bet} are shown in Table 12.11.

Table 12.11 Example 2: Calculation of SS_{Bet}

	M	$(M - M_G)$	$(M - M_G)^2$	$(M - M_G)^2n$
	Becomes:	$(M - 89.40)$	$(M - 89.40)^2$	$(M - 89.40)^2(5)$
Dorm	94.00	4.60	21.16	105.80
Home	93.00	3.60	12.96	64.80
Apartment	81.20	–8.20	67.24	336.20

$\Sigma(M - M_G)^2\, n = 506.80 = SS_{Bet}$

Alternatively, the same calculations can be presented as follows:

$$\begin{aligned} SS_{Bet} &= \Sigma[(M - M_G)^2 n] \\ &= [(94.00 - 89.40)^2 (5)] + [(93.00 - 89.40)^2 (5)] + [(81.20 - 89.40)^2 (5)] \\ &= [(4.60)^2 (5)] + [(3.60)^2 (5)] + [(-8.20)^2 (5)] \\ &= (21.16)(5) + (12.96)(5) + (67.24)(5) \\ &= 105.80 + 64.80 + 336.20 \\ &= 506.80 \end{aligned}$$

The value of SS_W can be found using the following equation:

$$SS_W = \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2$$

As these SS have been calculated in Table 12.8 calculation of SS_W is straightforward:

$$\begin{aligned} SS_W &= 46.00 + 46.00 + 14.80 \\ &= 106.80 \end{aligned}$$

We can use the following equation to check our calculations:

$$\begin{aligned} SS_T &= SS_{Bet} + SS_W \\ 613.60 &= 506.80 + 106.80 \\ 613.60 &= 613.60 \end{aligned}$$

The value of SS_W is *not* entered into our ANOVA table, as we now partition SS_W into $SS_{Subjects}$ and $SS_{Residual}$.

The $SS_{Subjects}$ is found by determining the square of the deviation of the mean of each subject from the grand mean, multiplying by the number of treatment levels, and then summing. Conceptually:

$$\begin{aligned} SS_{Subjects} &= [\Sigma(\frac{\Sigma \mathbf{X_{Subject}}}{\mathbf{k}} - \mathbf{M_G})^2\mathbf{k}] \\ &= [\Sigma(\mathbf{M_{Subject}} - \mathbf{M_G})^2\mathbf{k}] \end{aligned}$$

where the subject's total, $\Sigma X_{Subject}$, is obtained from Table 12.8, the grand mean, M_G, from Table 12.10 and k is the number of treatment levels, in this case 3.

These calculations are shown in Table 12.12.

Table 12.12 Example 2: Calculation of $SS_{Subjects}$

Subject	$\Sigma X_{Subject}$	$\mathbf{M_{Subject}} = \frac{\Sigma \mathbf{X_{Subject}}}{\mathbf{k}}$	$\mathbf{M_{Subject}} - \mathbf{M_G}$	$(\mathbf{M_{Subject}} - \mathbf{M_G})^2$	$(\mathbf{M_{Subject}} - \mathbf{M_G})^2\mathbf{k} = (\mathbf{M_{Subject}} - \mathbf{M_G})^2(\mathbf{3})$
1	278	92.67	3.27	10.69	32.07

2	272	90.67	1.27	1.61	4.83
3	272	90.67	1.27	1.61	4.83
4	262	87.33	–2.07	4.28	12.84
5	257	85.67	–3.73	13.91	41.73
			$\Sigma(\mathbf{M_{Subject}} – \mathbf{M_G}) \approx \mathbf{0}$		$\Sigma(\mathbf{M_{Subject}} – \mathbf{M_G})^{\mathbf{2}}(3)$
					$= 96.30 = \text{SS}_{\text{Subjects}}$

Alternatively, the same calculations can be presented as follows:

$$\begin{aligned}\text{SS}_{\text{Subjects}} &= [\Sigma(\mathbf{M_{Subject}} – \mathbf{M_G})^{\mathbf{2}}\mathbf{k}]\\ &= [(92.67 – 89.40^2](3) + [(90.67 – 89.40)^2](3) + [(90.67 – 89.40)^2](3) +\\ &\quad [(87.33 – 89.40)^2](3) + [(85.67 – 89.40)^2](3)\\ &= [(3.27)^2](3) + [(1.27)^2](3) + [(1.27)^2](3) + [(–2.07)^2](3) +\\ &\quad [(–3.73)^2](3)\\ &= (10.69)(3) + (1.61)(3) + (1.61)(3) + (4.28)(3) + (13.91)(3)\\ &= 32.07 + 4.83 + 4.83 + 12.84 + 41.73\\ &= 96.30\end{aligned}$$

The $\text{SS}_{\text{Subjects}}$ is the amount of variability that is being removed from the denominator of the F ratio by using a within-subjects design.

To determine the $\text{SS}_{\text{Residual}}$ we can subtract the value of $\text{SS}_{\text{Subjects}}$ from SS_{W} :

$$\text{SS}_{\text{W}} = \text{SS}_{\text{Subjects}} + \text{SS}_{\text{Residual}}$$

$$106.80 = 96.30 + \text{SS}_{\text{Residual}}$$

$$10.50 = \text{SS}_{\text{Residual}})$$

Calculating Degrees Of Freedom

Having calculated the needed SS, we now must calculate the degrees of freedom for between treatments, subjects, residual and total.

The degrees of freedom for between levels is equal to the number of treatment levels minus 1. Thus:

$$\text{df}_{\text{Bet}} = \text{k} – 1$$

where k is the number of treatment levels. In our example:

$$\begin{aligned}\text{df}_{\text{Bet}} &= 3 – 1\\ &= 2\end{aligned}$$

The degrees of freedom for subjects is equal to the number of subjects minus 1. Thus:

$$df_{Subjects} = n - 1$$

where n is the number of subjects. This equals:

$$\begin{aligned} df_{Subjects} &= 5 - 1 \\ &= 4 \end{aligned}$$

This is the number of degrees of freedom in the denominator that are removed from the analysis by using a within-subjects design compared to a between-subjects design.

To find the degrees of freedom for residual we subtract 1 from the total number of subjects and multiply this by the number of levels of the independent variable minus 1. Thus:

$$df_{Residual} = (n - 1)(k - 1)$$

where n is the number of subjects and k is the number of levels. In our example:

$$\begin{aligned} df_{Residual} &= (5 - 1)(3 - 1) \\ &= (4)(2) \\ &= 8 \end{aligned}$$

To find the degrees of freedom for total, we subtract 1 from the total number of data points. Thus:

$$df_T = N - 1$$

where N is the total number of data points. This equals:

$$\begin{aligned} df_T &= 15 - 1 \\ &= 14 \end{aligned}$$

As a check on our calculations:

$$df_T = df_{Bet} + df_{Subjects} + df_{Residual}$$

$$14 = 2 + 4 + 8$$

$$14 = 14$$

Calculating Mean Squares

The MS_{Bet} and the $MS_{Residual}$ are found by dividing the appropriate SS by its df. Thus:

$$\begin{aligned} MS_{Bet} &= \frac{\mathbf{SS_{Bet}}}{\mathbf{df_{Bet}}} \\ &= \frac{\mathbf{506.80}}{\mathbf{2}} \\ &= 253.40 \end{aligned} \qquad \begin{aligned} MS_{Residual} &= \frac{\mathbf{SS_{Residual}}}{\mathbf{df_{Residual}}} \\ &= \frac{\mathbf{10.50}}{\mathbf{8}} \\ &= 1.31 \end{aligned}$$

Calculating The F Ratio

Finally, we calculate the F ratio. Recall that for a within-subjects ANOVA the variability due to pre-existing subject differences was removed from the numerator of the F ratio due to the experimental design and is being mathematically removed from the denominator of the F ratio by removing the $SS_{Subjects}$. The denominator of the F ratio thus becomes $MS_{Residual}$. Therefore:

$$\begin{aligned} F &= \frac{\mathbf{MS_{Bet}}}{\mathbf{MS_{Residual}}} \\ &= \frac{\mathbf{253.40}}{\mathbf{1.31}} \\ &= 193.44 \end{aligned}$$

We have now completed the calculations for the ANOVA summary table (Table 12.13). We will later calculate the value in the final column of this table.

Table 12.13 Example 2: Completed Summary Table for the One-way Within-subjects ANOVA, with the Value for Partial Eta Squared (η_p^2)

Source of Variation	SS	df	MS	F	η_p^2
Between Treatments	506.80	2	253.40	193.44	0.98
Subjects	96.30	4			
Residual	10.50	8	1.31		
Total	613.60	14			

Interpreting The F Ratio

To determine whether this F ratio of 193.44 is statistically significant, in the F table (Appendix K, Table 4) we would locate the column corresponding to the degrees of freedom of our numerator and the row corresponding to the degrees of freedom of our denominator in the F ratio. From Table 12.13 we see that for our F this would be 2 and 8 degrees of freedom. At the intersection of this column and row in the F table for an α of .05 the critical value is 4.46. As our obtained F of 193.44 is greater than the critical value we reject the null hypothesis that housing choice does not affect academic success.

Conducting The Post Hoc Comparisons

Following a statistically significant within-subjects (repeated measures) ANOVA, a researcher can ask two substantially different questions when considering making post hoc comparisons. These questions are linked to the type of repeated measure that was utilized in the experiment. If the repeated measure consisted of a sequence on some dimension, such as different times, heights, distances or drug dosages, the researcher would most commonly focus upon the trend exhibited by the data – how the data changed across time, height, distance, or dosage – rather than upon which treatment levels differed. Consequently, they would probably employ what is called **trend analysis**, a topic that is beyond an introductory level statistics book.

Trend analysis – A statistical technique that attempts to define patterns in data.

Alternatively, the repeated measure might not consist of a sequence on some dimension. For example, a researcher might measure how test scores are affected by different drugs, or different types of exercise, or different types of background music. In these situations the IV doesn't consist of a natural sequence. Following a statistically significant one-way within-subjects ANOVA with data that are not sequential, the experimenter will focus upon the differences between treatment-level means. Of course, conducting all possible pairwise comparisons of these means is likely to lead to an increased probability of making Type I errors. To prevent this increased error we used the Tukey HSD as our post hoc test with the one-way between-subjects ANOVA. With the chi-square test of independence we employed the Bonferroni method. SPSS also uses the Bonferroni method to control for the increased likelihood of making Type I errors when conducting post hoc comparisons following a significant within-subjects ANOVA. Thus the discussion that follows, though differing in detail from what you learned previously, should nonetheless seem familiar.

While a significant F indicates that the IV had an effect, with three or more treatment levels it does not specify which treatment-level means differ. Remember, the data obtained from each treatment level of a repeated measures ANOVA are based on the same subjects. Therefore, we do not talk of a difference between samples or groups as we did when utilizing the one-way between-subjects ANOVA (Chapter 11) since there is now only one sample or group of subjects that is being repeatedly tested. Instead, we have differences between treatment levels.

It was noted in Chapter 11 that the total number of pairwise comparisons between sample (or treatment level) means is given by the equation:

$$\text{Number of pairwise comparisons} = \frac{\mathbf{k(k-1)}}{\mathbf{2}}$$

where k is the number of samples or treatment levels.

In our case, k equals 3, so there are [3(3 – 1)] / 2, which equals 3 pairwise comparisons. These 3 pairwise comparisons are between the mean of treatment level I and the mean of level II, the mean of level I and the mean of level III, and the mean of level II and the mean of level III. Any one, any two, or all three of these comparisons may be statistically significant. The significant F ratio simply indicates that at least one of the treatment level means is expected to differ from another. To specify which means differ we must once again conduct post hoc tests.

As was noted previously, following a significant one-way within-subjects ANOVA SPSS uses the Bonferroni method to control the Type I error rate when conducting all possible pairwise comparisons of treatment-level means. As you will recall, with the Bonferroni method the overall alpha rate you want to maintain is divided by the number of post hoc comparisons you will be making. The outcome is the per comparison alpha level. It is used in determining the critical value for each of the post hoc comparisons. Specifically, if we wish to have an overall Type I error rate (alpha level) of .05, and there are three post hoc comparisons, we divide .05 by 3 to obtain an alpha of .0167. Specifically:

Bonferroni method $= \frac{\textbf{\textit{alpha level to be maintained}}}{\textbf{\textit{number of post hoc comparisons}}} =$ critical value per comparison

In our case:

Bonferroni method $= \frac{\mathbf{.05}}{\mathbf{3}} = .0167$

Since the Bonferroni method is controlling the Type I error rate, any statistical test that utilizes interval or ratio data and that is appropriate for finding a difference with repeated measures could be used with our data. Table 12.1 indicates that either the one-way within-subjects ANOVA or the dependent samples t test could be utilized. Since SPSS uses the dependent samples t test, and the calculations are somewhat easier, we will also use this procedure for the post hoc comparisons. However, the decisions would be the same for either procedure.

For a post hoc comparison using the dependent samples t test, the df are defined as n – 1, where n is the number of subjects:

df = n – 1

In our case:

df = 5 – 1 = 4

The post hoc test is two-tailed. The critical value for an alpha of .0167 and 4 df is 3.96. (This value can be found using an online, t-test critical value calculator.) In other words, the value we calculate for each dependent samples t test must be greater than 3.96 in order to be considered statistically significant. It is important to note that this critical value of 3.96 is larger than the critical value would have been if the Bonferroni method had not been utilized. If the Bonferroni method had not been used, for an alpha of .05 and 4 df the critical value would have been 2.78. The increase in the size of the critical value is due to the Bonferroni method maintaining the overall error rate for the entire set of comparisons at .05 by making each of the post hoc comparisons more conservative.

The equation for the dependent samples t test, when the null hypothesis is that there isn't a difference ($\mu_{\mathbf{D}} = 0$), is:

$$t = \frac{\mathbf{M_D - \mu_D}}{\mathbf{s_{M_D}}} = \frac{\mathbf{M_D - 0}}{\mathbf{s_{M_D}}} = \frac{\mathbf{M_D}}{\mathbf{s_{M_D}}}$$

where:

$\mathbf{M_D}$ is the mean of the difference scores

$\mathbf{s_{M_D}}$ is the standard error of the mean difference

A single dependent samples t test is not difficult to obtain (refer to Chapter 10 and to the following Box). However, if you are conducting a series of these tests it will definitely save time to turn to SPSS instead of doing these calculations by hand. For illustration purposes, in the following calculations the value of $\mathbf{s_{M_D}}$, the standard error of the mean difference, is provided. The values of the treatment level means come from Table 12.8.

For our example we need to conduct 3 post hoc t tests:

For the comparison between the means of treatment level I and treatment level II:

$$\mathrm{t} = \frac{\mathbf{M_D}}{\mathbf{s_{M_D}}} = \frac{\mathbf{94.00-93.00}}{\mathbf{0.45}} = \frac{\mathbf{1.00}}{\mathbf{0.45}} = 2.22$$

For the comparison between the means of treatment level I and treatment level III:

$$\mathrm{t} = \frac{\mathbf{M_D}}{\mathbf{s_{M_D}}} = \frac{\mathbf{94.00-81.20}}{\mathbf{0.80}} = \frac{\mathbf{12.80}}{\mathbf{0.80}} = 16.00$$

For the comparison between the means of treatment level II and treatment level III:

$$\mathrm{t} = \frac{\mathbf{M_D}}{\mathbf{s_{M_D}}} = \frac{\mathbf{93.00-81.20}}{\mathbf{0.86}} = \frac{\mathbf{11.80}}{\mathbf{0.86}} = 13.72$$

It is important to note that when comparing these values of t to the critical value of 3.96, we would ignore the sign of each t test as this simply reflects the order the treatment level means were entered into the numerator for each calculation of t.

As our critical value is 3.96, the difference between the means of treatment level I and treatment level II, which results in a t value of 2.22 is not statistically significant. However, the differences between the means of level I and level III, which leads to a t value of 16.00, and between the means of level II and level III, which results in a t value of 13.72, are both statistically significant.

Box Showing Calculation of a Post Hoc Dependent t Test

Three post hoc comparisons of pairs of means were calculated for the data in Table 12.8. As discussed above, each of these comparisons involved calculation of a dependent t test. The outcome of each was then compared to the critical value found using the Bonferroni method. The calculations involved in the comparison between treatment level I and treatment level II are illustrated in Table 12.14. The calculations involved in the other two comparisons would be similar.

Table 12.14 Calculations for the Comparison between Treatment level I and Treatment level II

Subject	Dorm	Home	Difference Scores (D)	$(D - M_D)$	$(D - M_D)^2$
1	98	96	2	1.00	1.00

2	96	95	1	0.00	0.00
3	95	95	0	−1.00	1.00
4	91	91	0	−1.00	1.00
5	90	88	2	1.00	1.00
			$\sum D = 5$	$\sum(D - M_D) = 0$	$\sum(D - M_D)^2 = 4.00$

$M_D = \frac{\sum \mathbf{D}}{\mathbf{n_D}}$ where $\mathbf{n_D}$ = the number of difference

$= \frac{\mathbf{5}}{\mathbf{5}}$ scores, which is equal to the

$= 1.00$ number of *pairs* of scores

It is important to note that while there are two sets of scores in each post hoc comparison, there is only one sample of subjects and thus only one set of difference scores.

In our example, the null hypothesis is that the place of abode does not have an effect. The equation to determine t therefore is:

$$t = \frac{\mathbf{M_D - \mu_D}}{\mathbf{s_{M_D}}} = \frac{\mathbf{M_D - 0}}{\mathbf{s_{M_D}}} = \frac{\mathbf{M_D}}{\mathbf{s_{M_D}}}$$

The numerator of this equation, M_D, can be found by calculating $\sum D$ / $\mathbf{n_D}$, where $\mathbf{n_D}$ is equal to the number of difference scores, which is equal to the number of *pairs of scores*. As is indicated in Table 12.14, M_D for our example equals 5 / 5 or 1.00. It is important to recognize that this positive value of 1.00 indicates scores are *higher* when students live in the honors dorm than at home. The question we now need to address is whether this change of 1.00 point is statistically significant, and thus indicative of a reliable effect, or whether it should be considered to be the result of chance.

To find the standard error, $\mathbf{s_{M_D}}$, we note that $\mathbf{s_{M_D}} = s_D / \sqrt{\mathbf{n_D}}$. And, s_D, the estimate of the population standard deviation of a set of difference scores (which can alternatively be defined as the estimate of the population standard deviation of the differences between pairs of scores), is equal to:

$$s_D = \sqrt{\frac{\mathbf{\sum(D - M_D)^2}}{\mathbf{n_D - 1}}}$$

where $\mathbf{n_D}$ is equal to the number of difference scores, and is also equal to the number of *pairs* of scores.

Substituting from Table 12.14 we have:

$$s_D = \sqrt{\frac{\mathbf{4.00}}{\mathbf{5 - 1}}}$$

$$= \sqrt{\frac{\mathbf{4.00}}{\mathbf{4}}}$$

$$= \sqrt{1.00}$$

$$= 1.00$$

We can now determine the standard error, $\mathbf{s_{M_D}}$, by noting that:

$$\mathbf{s_{M_D}} = \frac{\mathbf{s_D}}{\sqrt{\mathbf{n_D}}}$$

$$= \frac{\mathbf{1.00}}{\sqrt{\mathbf{5}}}$$

$$= \frac{\mathbf{1.00}}{\mathbf{2.24}}$$

$$= 0.45$$

This is the denominator that we were seeking.

The value for t therefore becomes:

$$t = \frac{\mathbf{M_D}}{\mathbf{s_{M_D}}}$$

$$= \frac{\mathbf{1.00}}{\mathbf{0.45}}$$

$$= 2.22$$

This is the value of the dependent samples t test for the comparison of treatment level I and treatment level II that was utilized previously. An additional t test would need to be calculated for each of the other two post hoc comparisons.

Calculating The Effect Size

To ascertain the effect size for a one-way within-subjects ANOVA, SPSS calculates a **partial eta squared** ($\eta_{\mathbf{p}}^{\mathbf{2}}$).

Partial eta squared (η_{p}^{2}) – Measure of effect size calculated by SPSS for a within subjects ANOVA.

With a one-way within-subjects ANOVA an equation for the partial eta squared for treatment is:

$$\eta_{\mathbf{p}}^{\mathbf{2}} \text{ treatment} = \frac{\mathbf{SS_{Bet}}}{\mathbf{SS_T - SS_{Subjects}}}$$

This $\eta_{\mathbf{p}}^{\mathbf{2}}$ indicates the proportion of variance explained by the treatment, after removing the variability from the denominator due to pre-existing subject differences.

For our example:

$$\eta_{\mathbf{p}}^{\mathbf{2}} = \frac{\mathbf{506.80}}{\mathbf{613.60 - 96.30}}$$

$$= \frac{\mathbf{506.80}}{\mathbf{517.30}}$$

$$= .98 \text{ or } 98\%$$

This value is included in Table 12.13.

It is important to point out to the readers of this book that these data were created as an example and an effect size of .98 is much larger than is likely to be found in the real world. It indicates that 98% of the variability in the dependent variable (academic success) can be explained by the independent variable (housing choice). Only 2% of the variability, then, is due to all other factors, which is a very unlikely outcome.

Alternatively, we could calculate η^2 as was done with the one-way between-subjects ANOVA:

$$\eta^2 \text{ for treatment} = \frac{\mathbf{SS_{Bet}}}{\mathbf{SS_T}}$$

For our example:

$$\eta^2 = \frac{\mathbf{SS_{Bet}}}{\mathbf{SS_T}}$$

$$= \frac{\mathbf{506.80}}{\mathbf{613.60}}$$

$$= .83 \text{ or } 83\%$$

Note that the value of η^2 is less than the value of $\eta^2_{\mathbf{p}}$. This is because η^2 does not remove the variability from the denominator due to pre-existing subject differences while $\eta^2_{\mathbf{p}}$ does. There is not agreement on which is the better measure of effect size with a one-way within-subjects ANOVA. Most studies with a one-way within-subjects ANOVA probably report $\eta^2_{\mathbf{p}}$, but you could report either.

Reporting The Results Of A One-Way Within-Subjects ANOVA

In a paper, we would indicate the degrees of freedom, the F ratio that was obtained, as well as which pairwise comparisons were significant and the measure of effect size. Specifically, for these hypothetical data we would report (F(2,8) = 193.44, $p < .05$, $\eta^2_p = .98$) (You could report η^2 instead). This statement indicates that the treatment level means were found to differ and that our measure of effect size, $\eta^2_{\mathbf{p}}$, was found to equal .98. We would also note that post hoc comparisons utilizing the Bonferroni method test indicated that living in an off-campus apartment led to a decrease in exam scores compared to living in an on-campus honors dorm or living at home. However, no difference was found between living in an on-campus honors dorm and living at home. Later in this chapter we will use SPSS to analyze these data, and we then can make a more precise statement (F(2,8) = 192.46, $p < .001$, $\eta^2_p = .98$). The minor change in the value of the F ratio is due to rounding error in our calculations and SPSS provides a more precise p-value for the probability of our outcome. Also, note that the p-value is less than .001 which is also less than our α of .05, confirming that we would reject the null hypothesis. Finally, you should also report the results of Mauchly's test of sphericity which will be discussed in the SPSS section of this chapter.

Extension To Designs With Matched Subjects

This chapter has reviewed the one-way within-subjects ANOVA and the discussion has been limited to those designs in which the same sample of subjects are repeatedly tested. Thus the subjects being tested at each treatment level are not independent. The same statistical analysis is also employed in other research designs in which the subjects are related in some way. For instance, in order to begin an experiment with equivalent groups we could first ensure that the subjects were paired or matched on an important characteristic such as IQ. The matched sets of subjects would then be randomly assigned to the different treatment levels. The data from the resulting matched subjects design would also be analyzed with a one-way within-subjects ANOVA. In fact, repeatedly testing the same subjects can be thought of as the ultimate example of matching since we have essentially matched them on every possible preexisting characteristic.

Purpose And Limitations Of Using The One-way Within-subjects ANOVA

1. *Test for difference.* The null hypothesis is that the treatment does not have an effect. Therefore, if the null is correct any difference between the means of the treatment levels is due to chance. The alternative hypothesis is that the treatment does have an effect.
2. *Does not provide a measure of effect size.* The one-way within-subjects ANOVA, like the one-way between-subjects ANOVA, is a test of significance. It indicates whether or not an outcome is likely to have occurred by chance if the null hypothesis is correct. If the F test is significant a measure of effect size, such as eta squared (η^2) or partial eta squared (η_p^2), should then be calculated.
3. *Compares two or more treatment level means.* The one-way within-subjects ANOVA is appropriate to use when each subject is assigned to every treatment level, or when the subjects at each treatment level are matched on some variable.
4. *Does not indicate where the difference is.* With designs with more than two treatment levels, a significant F should be followed with a post hoc procedure. We have utilized a series of dependent t tests in order to identify which treatment level means differ. The Bonferroni method is used to control the Type I error rate.

Assumptions Of The One-way Within-subjects ANOVA

1. *Interval or ratio data.* The data are on an interval or a ratio scale of measurement.
2. *Random sample.* The subjects are drawn at random from a population.
3. *Independence within treatment levels.* The data within each treatment level are independent.
4. *Normally distributed populations.* The population at each treatment level is normally distributed. However, based upon the Central Limit Theorem, the one-way within-

subjects ANOVA will be accurate so long as the sample size is at least 30. If the sample size is less than 30, then it is important that the underlying populations be normally distributed. If you cannot collect a larger sample and have reason to believe that the assumption of normality may not have been met it is best to turn to an alternative test.

5. *Population variances are equal.* The population of scores corresponding to each treatment level have equal variances.
6. *No carryover effects.* With repeated measures, having received one treatment level does not affect a subject's response to another treatment level.

Effect Of Violating The Assumptions

As has been noted previously, the F test has been found to be robust – it often leads to accurate decisions even when an assumption is violated. However, when using a repeated measures design a researcher should be particularly concerned with carryover effects.

Conclusion

The one-way within-subjects ANOVA is a flexible, commonly employed statistical test to determine if treatment level means differ. Though somewhat tedious to calculate by hand, statistical packages such as SPSS make this a most useful statistical procedure.

As both the one-way between-subjects ANOVA and the one-way within-subjects ANOVA compare two or more treatment level means it is important to understand how they are related. Appendix M is designed to clarify how these ANOVAs are both similar and different.

It is also important to understand the advantages and disadvantages of each. As noted at the beginning of this chapter, the one-way within-subjects ANOVA design does not include preexisting subject differences in the numerator, and mathematically partitions out the preexisting subject differences from the denominator of the F ratio. As a result, the F ratio will likely be larger than if a one-way between-subjects ANOVA had been utilized. As a larger F ratio is more likely to be found to be statistically significant you might wonder why anyone would conduct a one-way between-subjects ANOVA. There are a number of reasons. First, you lose degrees of freedom with the one-way within-subjects ANOVA compared to the one-way between-subjects ANOVA. This is evident from the denominator of the F ratio, which in a one-way between-subjects ANOVA is MS_W but in a one-way within-subjects ANOVA it is $MS_{Residual}$. Each MS is associated with a degrees of freedom. For instance, in Table 12.13 $df_{Residual}$ is 8. However, if this had been a one-way between-subjects ANOVA the df_W would have been 12 (remember, $df_W = df_{Subjects} + df_{Residual}$). As an examination of the F table will indicate, this loss of 4 degrees of freedom results in a larger value of

F being needed in order to conclude that a difference is statistically significant. In other words, with a one-way within-subjects ANOVA the F ratio is likely to be larger than with a one-way between-subjects ANOVA, but as the degrees of freedom will definitely be smaller this F ratio will be compared to a larger critical value. Whether you gain more with the larger F than you lose due to the smaller degrees of freedom when using the one-way within-subjects ANOVA will depend upon the specific situation.

In addition, with a repeated measures design the experimenter needs to be concerned with carryover effects. This was mentioned previously in Chapter 10, but the basic idea is that with a repeated measures design it is being assumed that the effect of one treatment does not influence subsequent treatments. In some situations, however, this is unlikely to be the case. For instance, let's assume that in one condition subjects are assigned a physically demanding task, such as swimming 1,000 meters and in another condition they swim only 10 meters. The DV is the time needed to climb five flights of stairs after each swimming event. If there was not a very substantial rest period between the two swimming conditions it seems likely that swimming the long distance first would have a dramatic carryover effect on subjects' ability to subsequently climb stairs after swimming the short distance.

Another drawback to the repeated measures design is that subjects may not be willing to commit to repeated testing and thus will drop out of the study.

As a consequence of these limitations, the between-subjects design is much more commonly used than the repeated measures design and thus the one-way between-subjects ANOVA is more commonly used than the one-way within-subjects ANOVA. However, the repeated measures design, and thus the one-way within-subjects ANOVA, can be very useful, particularly if there are only a limited number of subjects available to be tested.

Final Thoughts On The Relationship Between The One-Way Within-Subjects ANOVA And The Dependent Samples t Test

Though the calculations for the dependent samples t test (reviewed in Chapter 10) and the one-way within-subjects ANOVA with two measures for each subject appear to be quite different, these tests are closely related. In fact, the outcome of the dependent samples t test and the outcome of the one-way within-subjects ANOVA are mathematically related in the same way that the independent samples t test and the one-way between-subjects ANOVA are related:

$F = t^2$

We utilized the same data using the dependent samples t test (Table 10.6) and the one-way within-subjects ANOVA (Table 12.3). Substituting the value of 0.51 found with the t test, we would have:

$F = 0.51^2 = 0.26$

which is the same value of F we found in this chapter except for minor rounding error (Table 12.7).

Thus, assuming you have interval or ratio data with two measures from each subject, or scores on two sets of matched subjects, you can consider conducting either a dependent samples t test or a one-way within-subjects ANOVA. Both will always lead to the same decision to retain or reject the null hypothesis. However, if you have three or more measures from each subject, or three or more sets of matched subjects, then you need to utilize the one-way within-subjects ANOVA.

Glossary Of Terms

Mauchly's test of sphericity – Statistical procedure utilized with SPSS to test the assumption of sphericity for a one-way within-subjects ANOVA.

One-way within-subjects ANOVA – An inferential procedure for comparing two or more means from related samples when there is one independent variable.

Partial eta squared (η_p^2) – Measure of effect size calculated by SPSS for a within subjects ANOVA.

Preexisting subject differences – Relatively stable subject characteristics. These differences between subjects are a form of error in an ANOVA. The variability due to these differences is removed in a one-way within-subjects ANOVA.

Residual error – Changeable subject characteristics. These differences between subjects are a form of error in an ANOVA. The variability due to these differences is not removed in a one-way within-subjects ANOVA.

Sphericity – Assumption of a within-subjects ANOVA that the variances of the sets of difference scores between treatment levels are equal. In a repeated measures ANOVA these differences would be based upon pairs of scores from each subject.

Sum of squares residual ($SS_{Residual}$) – In a one-way within-subjects ANOVA, the SS due to residual error.

Sum of squares subjects ($SS_{Subjects}$) – In a one-way within-subjects ANOVA, the SS due to preexisting subject differences.

Trend analysis – A statistical technique that attempts to define patterns in data.

Questions – Chapter 12

(Answers are provided in Appendix J.)

1. Compared to the one-way between-subjects ANOVA, the one-way within-subjects ANOVA _____.
   a. Reduces the size of the F ratio
   b. Removes the pre–existing subject differences
   c. Reduces residual error

d. Increases the size of the numerator of the F ratio

2. In the one-way within-subjects ANOVA, the SS_W is partitioned into _____.
   a. $SS_{Subjects}$ and $SS_{Residual}$
   b. SS_T and SS_{Bet}
   c. $SS_{Subjects}$ and SS_{Bet}
   d. SS_T and $SS_{Residual}$

3. M_G is the mean of _____.
   a. the scores in the largest group
   b. the scores in the first experimental condition
   c. the scores in the smallest group
   d. all of the scores

4. If the F is <u>not</u> significant, we _____.
   a. do not calculate the effect size, partial eta squared ($\eta_{\mathbf{p}}^{\mathbf{2}}$)
   b. can be absolutely certain that our independent variable did not have an effect
   c. should consider conducting the study again, but this time with fewer subjects
   d. should then conduct a post hoc test

5. If the F is significant, and we have more than 2 treatment levels for each subject, we would _____.
   a. calculate the effect size, partial eta squared ($\eta_{\mathbf{p}}^{\mathbf{2}}$)
   b. conduct post-hoc comparisons
   c. both of the above
   d. none of the above

6. Compared to the one-way between-subjects ANOVA, with a one-way within-subjects ANOVA _____.
   a. There is a loss of df
   b. A larger F value is needed
   c. There are fewer calculations
   d. Both ‘a’ and ‘b’, but not ‘c’

7. Which is more commonly utilized, the one-way between-subjects ANOVA or the one-way within-subjects ANOVA?
   a. One-way between-subjects ANOVA
   b. One-way within-subjects ANOVA
   c. Both are used approximately equally often

8. Eye color is an example of _____ and catching a cold is an example of _____.
   a. Residual error; residual error
   b. Residual error; pre-existing subject differences
   c. Pre-existing subject differences; pre-existing subject differences
   d. Pre-existing subject differences; residual error

9. The F ratio for a one-way within-subjects ANOVA is equal to _____.
   a. MS_{Bet} / MS_W
   b. MS_{Bet} / $MS_{Subjects}$
   c. MS_{Bet} / $MS_{Residual}$
   d. MS_{Bet} / SS_{Total}

10. If your design has 10 treatment levels, how many post hoc pairwise comparisons would there be?
   a. 3
   b. 10
   c. 33
   d. 45

11. If in a paper you read ($F(3, 9) = 61.44$, $p < .05$), what does the number 3 refer to?
   a. df_{Bet}
   b. $df_{Subjects}$
   c. $df_{Residual}$
   d. df_T

12. With a one-way within-subjects ANOVA, partial eta squared is the _____.
   a. proportion of variance explained by the subjects
   b. proportion of variance explained by the treatment
   c. proportion of variance explained by the error
   d. proportion of variance explained by the residual

For questions 13 – 16 we are going to use the same data as in Chapter 11 (Questions 15 – 19) except that we now assume there are only a total of 6 students and each student took different versions of the exam in the quiet, moderately noisy and noisy environments. Compare each of your answers to the answer you calculated in Chapter 11.

		Level of Background Noise		
		Quiet	Moderate	Noisy
	1	9	7	6
	2	10	9	8
Student	3	8	8	10
	4	13	13	7
	5	12	11	11
	6	14	12	12

13. What is the SS for the level of background noise (SS_{Bet})?
   a. 2.0
   b. 12.0
   c. 2.43
   d. 6.0

14. What is the df for the level of background noise?
   a. 2
   b. 12
   c. 2.43
   d. 6

15. What is the MS for the level of background noise?
   a. 2.0
   b. 12.0
   c. 2.43
   d. 6.0

16. What is the value of F ratio?
   a. 2.0

b. 12.0
c. 2.43
d. 6.0

Problems 17 – 20 utilize SPSS.

Using SPSS With The One-way Within-subjects ANOVA

To Begin SPSS

Step 1 Activate the program, close the central window, and click on the **Variable View** option at the bottom left of the window.

Step 2 Click on the first empty cell under the column heading 'Name'. You now type the name of the first variable for which you have data. We are going to utilize the same data and labels as were previously employed in Table 12.8. These data dealt with the effect of living situation in college students. Type 'Dorm' in the first empty cell under 'Name'.

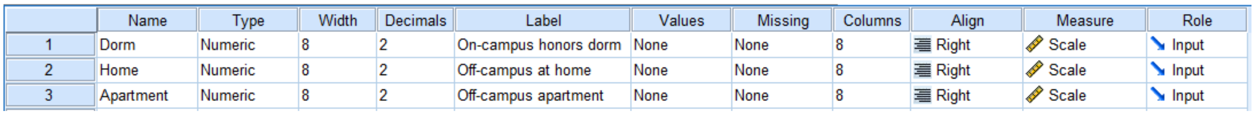
Step 3 Click on the first empty 'cell' under the column heading 'Label'. In this cell you can type a more extensive description of your variable. In our case, type 'On-campus honors dorm'. Note that in order to see the entire label you may need to expand the size of this cell by placing your cursor on the right border of the Label heading and moving to the right.

Step 4 Check the first 'cell' under the column heading 'Measure'. As we are dealing with exam scores be certain that 'Scale' is present.

Step 5 Repeat Steps 2 – 4 except that you type 'Home' in the first empty cell under 'Name' and 'Off-campus at home' for the label. Finally, select 'Scale' in the column under the column heading 'Measure' as we have ratio data.

Step 6 Repeat Steps 2 – 4 except that you type 'Apartment' in the first empty cell under 'Name' and 'Off-campus apartment' for the label. Finally, select 'Scale' in the column under the column heading 'Measure' as we have ratio data. The result is shown in Figure 12.1.

Figure 12.1 Variable View Window



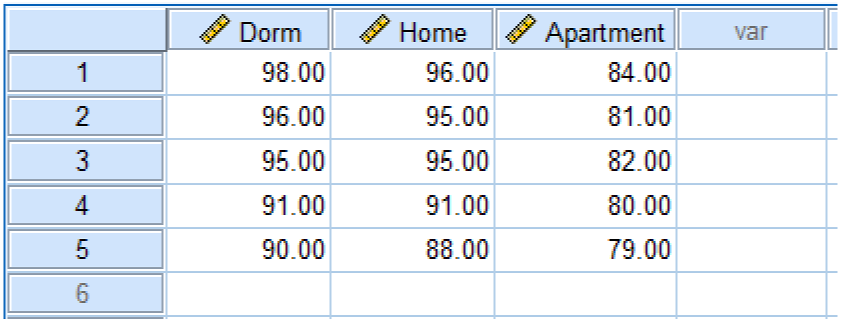
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Dorm	Numeric	8	2	On-campus honors dorm	None	None	8	Right	Scale	Input
2	Home	Numeric	8	2	Off-campus at home	None	None	8	Right	Scale	Input
3	Apartment	Numeric	8	2	Off-campus apartment	None	None	8	Right	Scale	Input

To Enter Data In SPSS

Step 7 Click on the '**Data View**' option at the lower left corner of the Variable View window. The variables 'Dorm', 'Home' and 'Apartment' will be evident.

Step 8 Our data consist of three exam scores for each of five subjects. For each subject enter the hypothetical exam scores in the appropriate row and columns, as is shown in Figure 12.2.

Figure 12.2 Entering Data



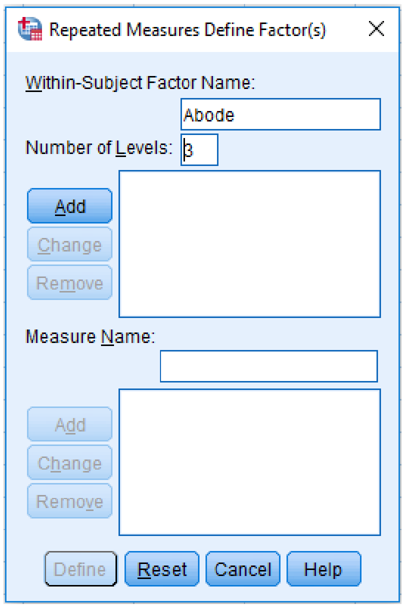
	Dorm	Home	Apartment	var
1	98.00	96.00	84.00	
2	96.00	95.00	81.00	
3	95.00	95.00	82.00	
4	91.00	91.00	80.00	
5	90.00	88.00	79.00	
6				

To Conduct A One-way Within-subjects ANOVA

Step 9 Click the cursor on '**Analyze**' along the row of SPSS commands above the data you entered, then move to '**General Linear Model**', then click on '**Repeated Measures**'.

Step 10 A new window will appear. This asks for the 'Within–Subject Factor Name'. In our case, Abode would be an appropriate name. This is indicated by typing 'Abode' in the upper box and '3' for the 'Number of Levels' as we have three living situations (Figure 12.3). Then click on '**Add**'.

Figure 12.3 Repeated Measures Window

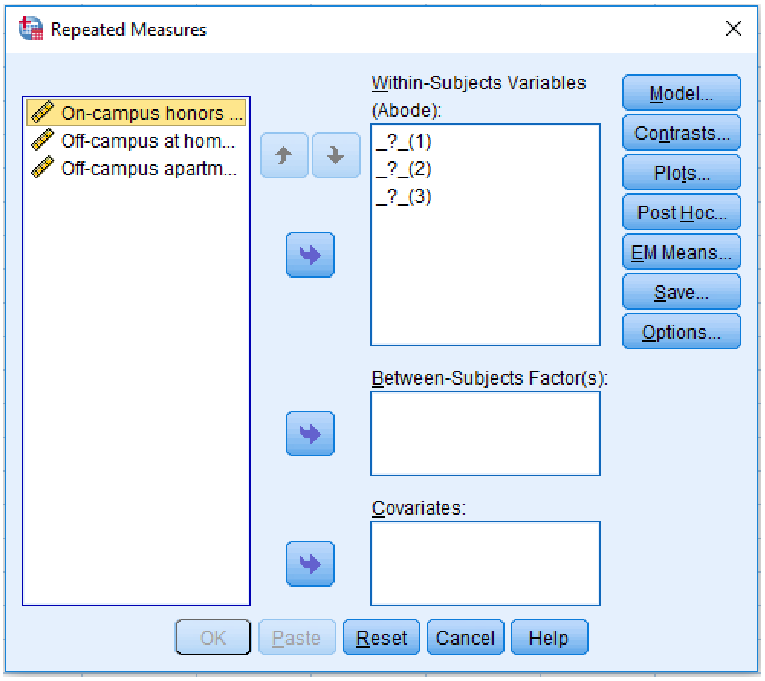


Within-Subject Factor Name: Abode

Number of Levels: 3

Step 11 Click '**Define**' and a new window will appear (Figure 12.4). Move each of the labels on the left to the box on the right by clicking on the appropriate label and then on the top arrow pointing to the right. This assigns a variable to each of the levels you have noted in Step 10 (in our case there are three levels). The result will appear as is shown in Figure 12.5.

Figure 12.4 Defining the Repeated Measures Variable



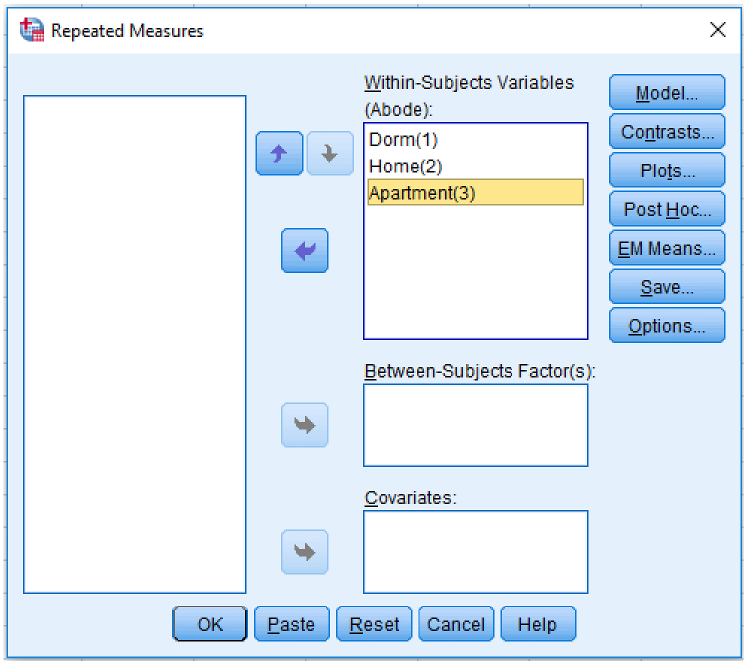
On-campus honors ...
Off-campus at hom...
Off-campus apartm...

Within-Subjects Variables (Abode):
_?_(1)
_?_(2)
_?_(3)

Between-Subjects Factor(s):

Covariates:

Figure 12.5 Continuing to Define the Repeated Measures Variable



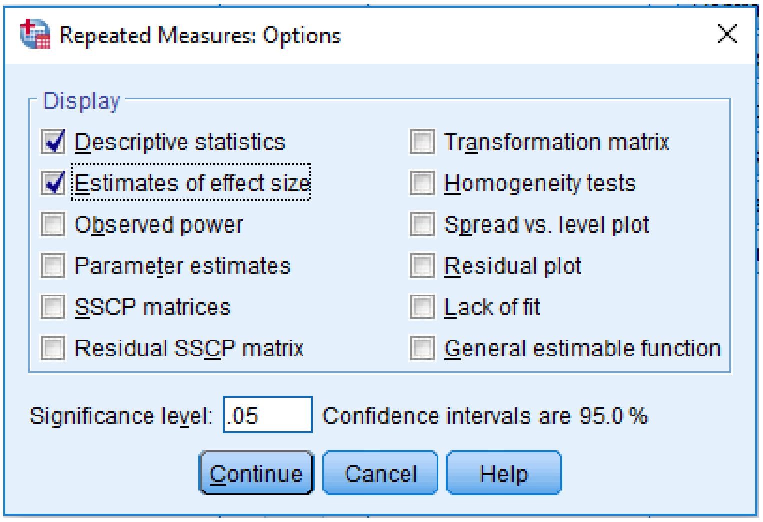
Within-Subjects Variables (Abode):
Dorm(1)
Home(2)
Apartment(3)

Between-Subjects Factor(s):

Covariates:

Step 12 Now click on '**Options**' which is located on the column on the right. A new window will appear. If you click on the boxes in front of '**Descriptive statistics**' and '**Estimates of effect size**' (Figure 12.6) SPSS will later generate a useful summary of the data. Click '**Continue**'.

Figure 12.6 Specifying Descriptives and Effect Size



Repeated Measures: Options

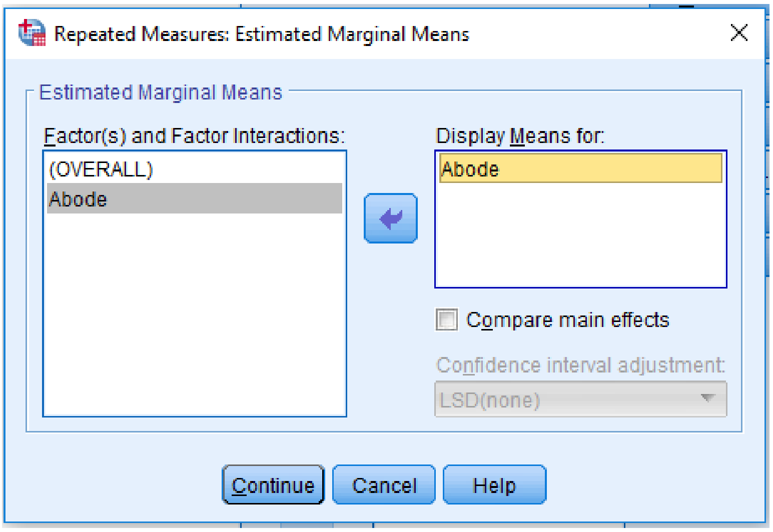
Display

☑ Descriptive statistics
☑ Estimates of effect size
□ Observed power
□ Parameter estimates
□ SSCP matrices
□ Residual SSCP matrix
□ Transformation matrix
□ Homogeneity tests
□ Spread vs. level plot
□ Residual plot
□ Lack of fit
□ General estimable function

Significance level: .05 Confidence intervals are 95.0 %

Step 13 Now click on 'Abode' and then the central arrow. Abode will move into the box labelled 'Display Means for:' (Figure 12.7).

Figure 12.7 Specifying the Post Hoc



Repeated Measures: Estimated Marginal Means

Estimated Marginal Means

Factor(s) and Factor Interactions:
(OVERALL)
Abode

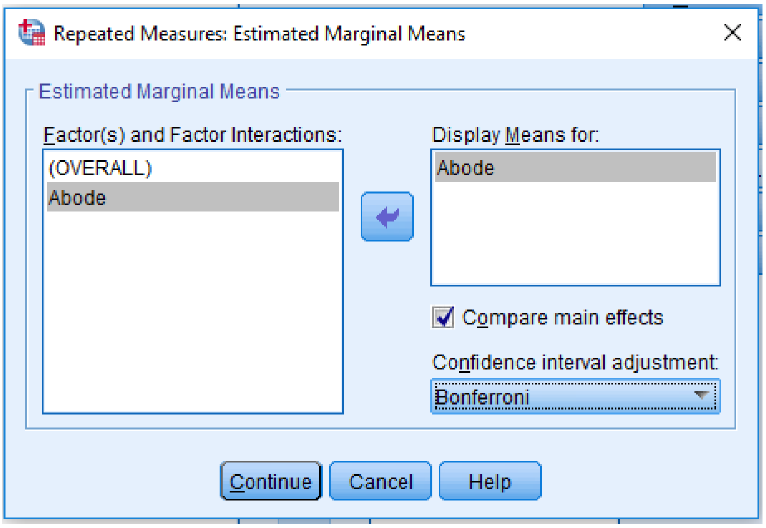
Display Means for:
Abode

□ Compare main effects

Confidence interval adjustment:
LSD(none)

Step 14 Click on the box in front of '**Compare main effects**'. Now click on the drop down menu that says 'LSD(none)' and select 'Bonferroni' (Figure 12.8). Click on '**Continue**'.

Figure 12.8 Specifying the Bonferroni as the Post Hoc



Repeated Measures: Estimated Marginal Means

Estimated Marginal Means

Factor(s) and Factor Interactions:

(OVERALL)

Abode

Display Means for:

Abode

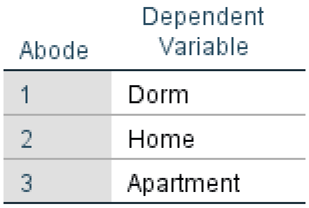
☑ Compare main effects

Confidence interval adjustment:

Bonferroni

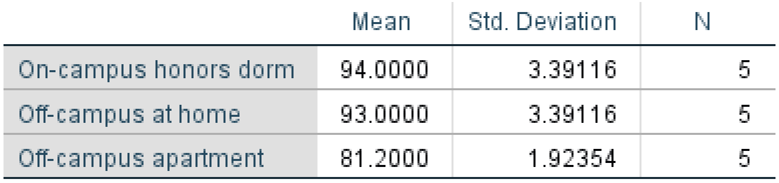
Step 15 You will be returned to Figure 12.5. Click '**OK**' and SPSS will conduct the one-way within-subjects ANOVA. The printout is quite complex. The parts of the output that we are interested in have the headings 'Within-Subjects Factors' (Table 12.15), 'Descriptive Statistics' (Table 12.16). 'Mauchly's Test of Sphericity' (Table 12.17), 'Tests of Within-Subjects Effects' (Table 12.18) and 'Pairwise Comparisons (Table 12.19).

Table 12.15 SPSS Output; Within-Subjects Factors



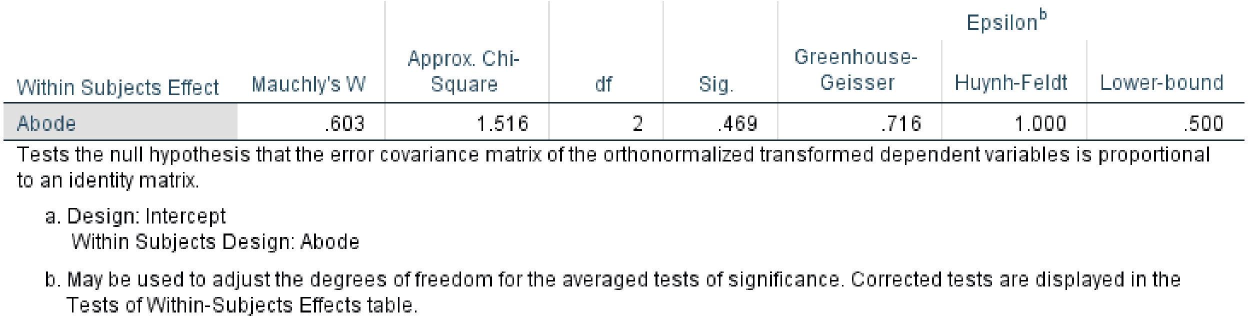
Abode	Dependent Variable
1	Dorm
2	Home
3	Apartment

Table 12.16 SPSS Output; Descriptive Statistics



	Mean	Std. Deviation	N
On-campus honors dorm	94.0000	3.39116	5
Off-campus at home	93.0000	3.39116	5
Off-campus apartment	81.2000	1.92354	5

Table 12.17 SPSS Output; Mauchly's Test of Sphericity



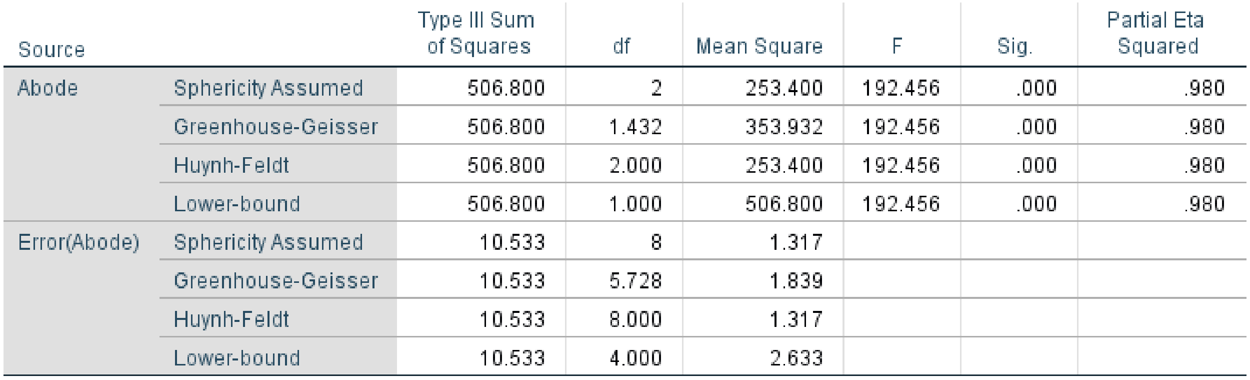
Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Abode	.603	1.516	2	.469	.716	1.000	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept
Within Subjects Design: Abode

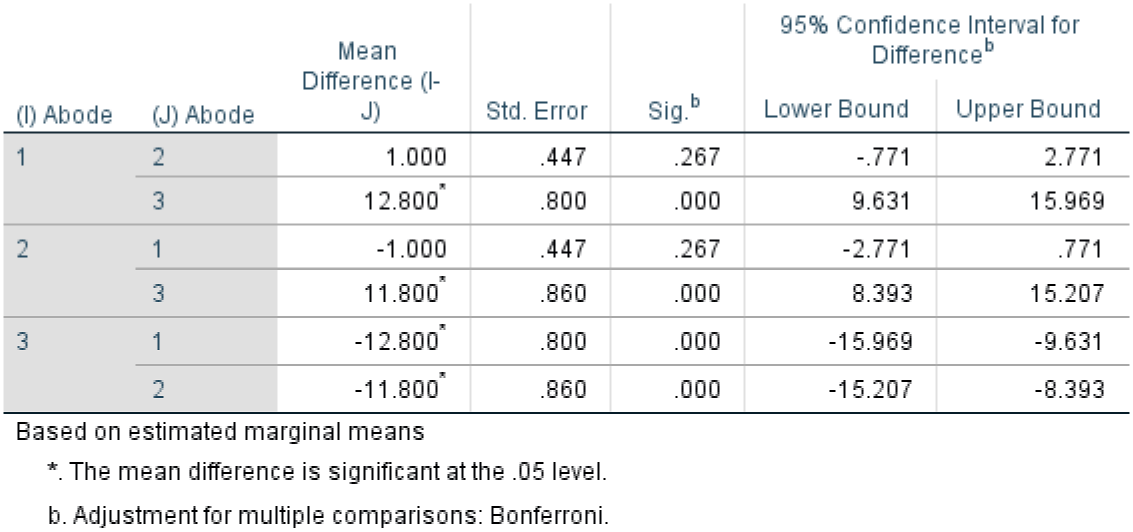
b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

Table 12.18 SPSS Output; Tests of Within-Subjects Effects



Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Abode	Sphericity Assumed	506.800	2	253.400	192.456	.000	.980
	Greenhouse-Geisser	506.800	1.432	353.932	192.456	.000	.980
	Huynh-Feldt	506.800	2.000	253.400	192.456	.000	.980
	Lower-bound	506.800	1.000	506.800	192.456	.000	.980
Error(Abode)	Sphericity Assumed	10.533	8	1.317			
	Greenhouse-Geisser	10.533	5.728	1.839			
	Huynh-Feldt	10.533	8.000	1.317			
	Lower-bound	10.533	4.000	2.633			

Table 12.19 SPSS Output; Pairwise Comparisons



(I) Abode	(J) Abode	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
1	2	1.000	.447	.267	-.771	2.771
	3	12.800*	.800	.000	9.631	15.969
2	1	-1.000	.447	.267	-2.771	.771
	3	11.800*	.860	.000	8.393	15.207
3	1	-12.800*	.800	.000	-15.969	-9.631
	2	-11.800*	.860	.000	-15.207	-8.393

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.

Table 12.15 identifies the levels of the IV, and thus what condition the data (DV) refer to. Table 12.16 provides the mean exam score and standard deviation for each of the three levels of abode. Table 12.17 lists the result of Mauchly's Test of Sphericity. Recall that sphericity is an assumption of a within-subjects ANOVA. Briefly, it is that the differences between scores at pairs of levels have equal variances (Table 12.9). Table 12.17 indicates the value of Mauchly's Test of Sphericity is .603 and the p-value (Sig.) is .469. To reject the assumption of sphericity, the p-value would need to be less than .05. Since .469 is greater than .05, the assumption of sphericity is retained and we can proceed to the ANOVA. We can ignore the remainder of this table. Table 12.18 provides the ANOVA table summary. Refer to the two rows labeled 'Sphericity Assumed' since Mauchly's test did not lead us to reject this assumption. (If Mauchly's test was significant, refer to a more advanced statistics text.) These rows begin with the labels 'Abode' (what we called 'Between Treatments') and 'Error(Abode) (what we called 'Residual'). What you will see is the same result for the ANOVA, except for our minor rounding error, as we previously found with hand calculations (Table 12.13) and the same value for partial eta squared that we found previously. The last table listed (Table 12.19) summarizes the outcome of the post hoc dependent t tests with the Bonferroni method. The presence of an asterisk in the column 'Mean Difference' indicates that the hypothetical exam scores for students when they lived in Abode 3 (Apartment) were significantly less than the scores for when they lived in Abode 1 (Honors Dorm) or Abode 2 (At Home). Further, the lack of an asterisk in the column 'Mean Difference' indicates that there was not a significant difference between the exam scores for Abode 1 and Abode 2.

Finally, just as was the case with the dependent samples t test, note that Levene's test for homogeneity of variances is not utilized with a within-subjects ANOVA.

Step 16 Exit SPSS. There is no need to save the output or the data.

To confirm that you understand how to use SPSS, I suggest you redo the ANOVA dealing with level of background noise (Questions 13 – 16) to check your answers.

SPSS Problems – Chapter 12

Problems 17 – 20 are based upon the same data that were used for questions 13 – 16 except that we now want to determine the effect of adding a constant (in this case 10) to every score in the noisy condition. (Compare your answers to the answers for questions 13 – 16 in this chapter and the answers in Chapter 11 for questions 25 and 26 when a between-subjects ANOVA was utilized.)

		Level of Background Noise		
		Quiet	Moderate	Noisy
	1	9	7	16
	2	10	9	18
Student	3	8	8	20
	4	13	13	17
	5	12	11	21
	6	14	12	22

17. What is the SS for the level of background noise (SS_{Bet})?
a. 2.0

b. 12.0
c. 292.0
d. 6.0

18. What is the df for the level of background noise?
a. 2
b. 12
c. 2.43
d. 6

19. What is the MS for the level of background noise?
a. 2.0
b. 12.0
c. 2.43
d. 146.0

20. What is the value of F ratio?
a. 2.659
b. 14.443
c. 26.500
d. 59.189

21. Based upon the post hoc analysis, which treatment levels differ?
a. None of the treatment levels differ
b. Quiet differs from Moderate, but not from Noisy
c. Moderate doesn’t differ from Quiet or Noisy
d. Noisy differs from both Quiet and from Moderate

Chapter 13
Finding Differences with Interval/Ratio Data – V: The Two-way Between-subjects ANOVA

"Oh, fancies that might be, oh facts that are!"

Robert Browning

Introduction

In Chapter 11 we reviewed the one-way between-subjects ANOVA. It is among the most commonly used of all statistical procedures. In Chapter 12 we reviewed the one-way within-subjects ANOVA, a useful alternative to a between-subjects design. Nevertheless, both of these ANOVAs are limited because they examine the effect of only one IV. In the real world we are simultaneously affected by numerous variables. For instance, your comprehension of this chapter will depend upon many factors including how much sleep you got last night, whether you are under time pressure, how noisy the background is, and your understanding of previous chapters, to name just a few. In this chapter we will learn that an ANOVA can be utilized when there is more than one IV. We will only be discussing the situation where there are two IVs. Though an ANOVA can maintain the experimentwise error rate while simultaneously dealing with an unlimited number of IVs the analysis quickly becomes difficult to interpret.

In Chapter 11 we learned that when dealing with an ANOVA each IV is called a factor. Thus, the single-factor or one-way ANOVA has only one IV and, if it is a between-subjects design each subject experiences only one level of the IV. An ANOVA with more than one factor is called a factorial ANOVA. To describe a factorial ANOVA the number of levels of each IV is specified. Thus, if there are two IVs, each with two levels, this would be a 2 X 2 ANOVA (this is read, "two by two ANOVA"). If there were two IVs, one with two levels and the other with three levels, this would be a 2 X 3 ANOVA. If there were three IVs, one with two levels and two with three levels, this would be a 2 X 3 X 3 ANOVA. In this chapter we will only be dealing with designs with two IVs and where there are no repeated measures or matched subjects. We will, accordingly, be studying what is called the **two-way between-subjects ANOVA** or two-factor between-subjects ANOVA. This procedure is underlined in Table 13.1.

Two-way between-subjects ANOVA – An inferential procedure for comparing means from independent samples when there are two independent variables.

Table 13.1 Overview of Inferential Statistical Procedures For Finding if there is a Difference

	Type of Data			
	Nominal (Frequency)		Ordinal (Ranked)	Interval/Ratio (Continuous Measure)
Research Design		Research Design		
One Variable With At Least Two Outcomes	Goodness-of-fit Chi-Square	One IV With One Sample		One-sample z Test or One-sample t Test
		One IV With Two Or More Independent Samples	*Kruskal–Wallis H*	One-way Between–Subjects ANOVA (Only two independent samples, Independent Samples t Test)
		One IV With One Sample Having Two Or More Repeated Measures		One-way Within–Subjects ANOVA (Only two repeated measures, Dependent Samples t Test)
Two Variables, Each With At Least Two Outcomes	Chi-Square Test of Independence	Two IV Each With Two Or More Independent Samples		Two-way Between–Subjects ANOVA

The Italicized procedure is reviewed in Appendix A

Main Effects And Interaction

An example of a 2 X 3 between-subjects ANOVA is shown in Table 13.2. The two IVs are gender (2 levels; men or women) and academic major (3 levels; arts, sciences, or other). The DV is grade point average. Which IV is designated Factor A and which is Factor B is arbitrary. In Table 13.2 gender is Factor A (rows) and academic major is Factor B (columns). Each of the six combinations of the levels of Factor A and Factor B is called a **cell**, and each cell is numbered as you would read a page. Note that in a between-subjects factorial ANOVA each subject is assigned to a single cell and thus experiences only one combination of treatment levels.

Cell – A particular combination of treatment levels in a Factorial ANOVA.

Table 13.2 Illustration of a 2 X 3 Between-Subjects ANOVA

Factor B

Academic Major

		Arts	Sciences	Other
Factor A	Men	Cell 1	Cell 2	Cell 3

Gender	Women	Cell 4	Cell 5	Cell 6

With a two-way ANOVA we can examine the effect of each of the two IVs separately. These are called **main effects**. The two-way ANOVA is, therefore, somewhat like simultaneously conducting two, one-way ANOVAs. In addition, however, with a two-way ANOVA we can also examine how the two IVs interact with each other. For instance, it has been reported in a number of publications that childhood maltreatment leads to antisocial adult behavior. This effect is sometimes summarized by saying that childhood abuse runs in families. In other words, it has been accepted that there is a relationship between the level of one variable (childhood abuse) and the magnitude of another variable (degree of antisocial adult behavior). Caspi et al. (2002) reexamined the long-term effects of childhood maltreatment. This study differed from the previous research by including an additional variable, the presence or absence in the subjects of a gene encoding the monoamine oxidase A (MAOA) enzyme. It was found that those men with low MAOA levels were much more likely to have a record of antisocial behavior, but only if they had been abused as children. The men with high MAOA levels were not antisocial even if they had been abused as children. This study suggests, therefore, that our previous interpretation was only partially correct. While there is a link between childhood abuse and adult antisocial behavior, this relationship appears to be dependent upon the individual's genetic makeup. In other words, Caspi et al. (2002) found that antisocial adult behavior is only enhanced when two factors occur together. Neither, by itself, is sufficient to lead to elevated rates of adult antisocial behavior. This dependency of an effect upon a combination of factors is called an **interaction**.

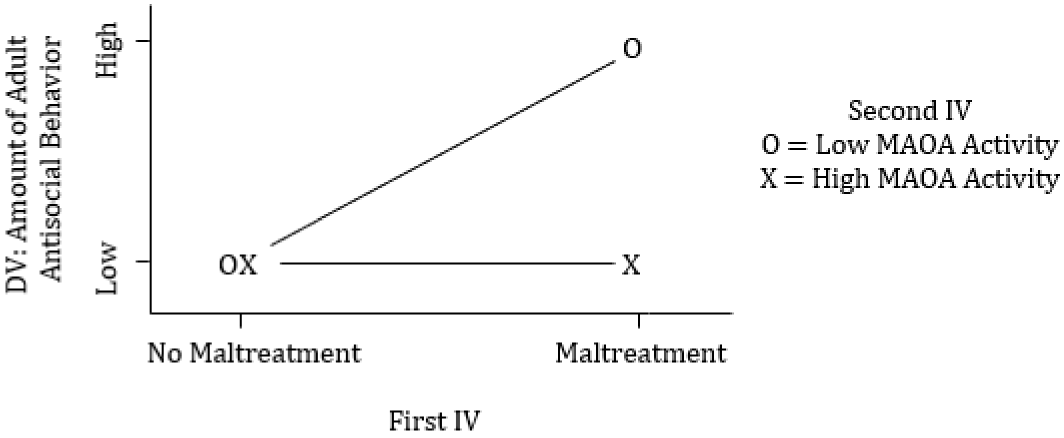
Main effect – With a factorial ANOVA, another term for an independent variable or factor.

Interaction – A change in the dependent variable that is due to the presence of a particular combination of independent variables.

We are all familiar with the concept of an interaction. For instance, physicians warn against taking particular combinations of medications. Though each medication may be helpful by itself, the combination may definitely not be. It is also widely known that the combination of two useful household cleaners, ammonia and chlorine bleach, will lead to the production of chlorine gas which is very dangerous. Finally, the author of this book likes to eat pickles and also likes ice cream, but the combination of pickles and ice cream does not sound appealing. Thus, in an interaction the combined effect of two factors is not simply the sum of the effects of the two factors alone.

One of the most useful techniques to assist in interpreting interactions is to graph the outcome. Returning to the Caspi et al. (2002) study that examined the long-term effects of childhood maltreatment, we could assign one of our IVs, childhood maltreatment, to the X axis, the amount of adult antisocial behavior, the DV, to the Y axis, and then plot the outcome for the two levels of the second IV, genetic makeup. The result is illustrated in Figure 13.1.

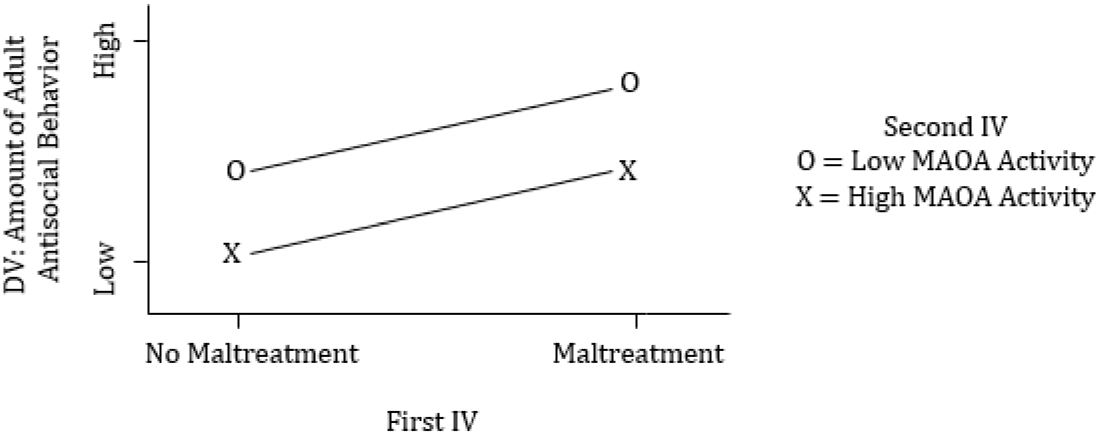
Figure 13.1 An Example of an Interaction



The advantage of a graph is that the interaction is evident at a glance. The amount of adult antisocial behavior is only increased if there was a history of maltreatment and if the individual had low MAOA activity.

Now let us compare what the graph would have looked like if there had not been an interaction. Specifically, if the outcome had been that both IVs had an effect but there was no interaction between the two factors, we might find an outcome as in Figure 13.2.

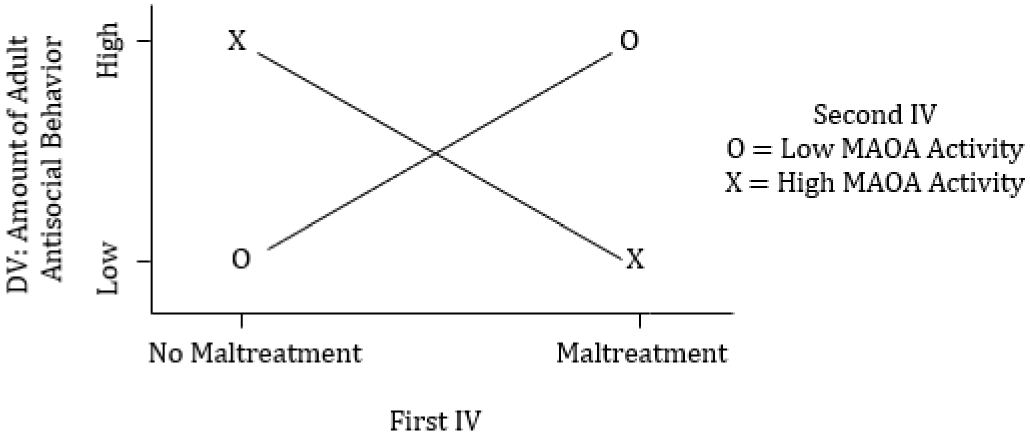
Figure 13.2 An Example of Two Main Effects but no Interaction



In this example, childhood maltreatment would have increased adult antisocial behavior and low MAOA activity would also have been associated with higher adult antisocial behavior. Thus, in Figure 13.2 both IVs had significant effects. Consequently, another way to describe our outcome would be to say that there were two significant main effects. Having found a significant main effect does not indicate whether, or not, there is also a significant interaction. In the example illustrated in Figure 13.2 there is not an interaction as no particular combination of the IVs causes a unique change in the DV. In other words, any outcome is explained by simply adding the separate effects of each IV.

With a two-way ANOVA, there are two IVs or factors, but three F ratios are calculated. One F ratio is calculated for each of the two possible main effects, and a third F ratio is calculated to determine if there is an interaction. None, or any combination of these three F ratios can be significant. In other words, neither, one, or both of the main effects might be found to be significant, and the F ratio for the interaction could be significant regardless whether any main effect was found to be significant. For instance, returning to our example of the effects of maltreatment, Figure 13.3 would be an example of a significant interaction though neither main effect is significant.

Figure 13.3 An Example of a Significant Interaction but no Significant Main Effects



If the results had been as shown in Figure 13.3, there would not be a main effect for the maltreatment IV since the overall amount of adult antisocial behavior is the same regardless of whether the child was maltreated or not. Similarly, with the results portrayed in Figure 13.3 there is not a main effect for the IV of MAOA activity since the overall, or mean, amount of adult antisocial behavior is the same regardless of the subject's MAOA activity. However, there is an interaction. As drawn, the results would indicate that there would be an increase in adult antisocial behavior either with no maltreatment and high MAOA activity, or with maltreatment and low MAOA activity. When an interaction is found to be significant, it, not the main effects, becomes the center of our attention. We return now to Figure 13.1, which is a representation of the results actually found by Caspi et al. (2002). The interaction suggests that, in addition to trying to reduce the overall level of maltreatment, to counter adult antisocial behavior we might consider a special focus upon male children who exhibit low MAOA activity.

The Logic Of The Two-Way Between-Subjects ANOVA

The two-way between-subjects ANOVA can be understood as an extension of the one-way between-subjects ANOVA that was covered in Chapter 11. With the one-way between-subjects ANOVA we calculate two estimates of the population variance (σ_X^2). The within-groups estimate is called the mean square within (MS_W). The MS_W is the estimate of σ_X^2 obtained by pooling the variability *within* each of the experimental groups. Since each subject within an experimental group receives the same level of the treatment, this variability is not a result of the IV and, instead, is due to other sources of variability, which we collectively call error. The other estimate of σ_X^2 in a one-way between-subjects ANOVA is called the mean square between (MS_{Bet}). The MS_{Bet} is the estimate of σ_X^2 based on the variability *between* the groups. Each experimental group receives a different level of the treatment. Thus, this variability is the result of the IV as well as what we are calling error. As a result, with the one-way between-subjects ANOVA we have two methods for estimating σ_X^2. And if there is no treatment effect, these two estimates of σ_X^2 are expected to be approximately the same. However, if the IV had an effect, the two estimates of σ_X^2 may differ substantially.

With the one-way between-subjects ANOVA we calculate one F ratio. If there is no treatment effect, the ratio of MS_{Bet} / MS_W should be approximately 1.00. If there is a treatment effect, the F ratio will be greater than 1.00.

With a two-way between-subjects ANOVA there is still a within-groups estimate of σ_X^2, the MS_W. However, with a two-way between-subjects ANOVA the MS_W is the estimate of σ_X^2 obtained by pooling the variances derived from each score's deviation from its *cell mean*. Since all subjects within a cell are treated similarly (they receive the same combination of levels of the two IVs) this variability is not a result of receiving different treatments and, instead, is due to other sources of variability, which we again collectively call error. The other estimate of σ_X^2 in a *one-way* between-subjects ANOVA is called the mean square between (MS_{Bet}). However, with a *two-way* between-subjects ANOVA the SS_{Bet}, which is the basis for the MS_{Bet}, is partitioned to create an estimate of σ_X^2 from each of the two IVs as well as an estimate from the interaction between the two IVs. With ANOVAs you will recall that IVs are called factors. Therefore, it is customary to say that we partition the SS_{Bet} that is used to create the MS_{Bet} estimate of σ_X^2 into the variability accounted for by Factor A, the variability accounted for by Factor B and, finally, the variability accounted for by the interaction between Factor A and Factor B (after removing any unique effects of Factor A and Factor B). The variability accounted for by Factor A is used to create a new estimate of σ_X^2 called the mean square for Factor A (MS_A). The variability accounted for by Factor B is used to create a second estimate of σ_X^2 called the mean square for Factor B (MS_B). (It is important to note that MS_B

is not the same as MS_{Bet}.) Finally, the variability accounted for by the interaction of Factor A and Factor B is used to create a third estimate of $\sigma_X{}^2$ called the mean square for the interaction of Factors A and B (MS_{AXB}). (Note that the interaction is represented as A X B, and MS_{AXB} is read as "mean square A times B.) We will, therefore, be calculating three F ratios with a two-way between-subjects ANOVA:

F_A, the main effect of Factor A, = MS_A / MS_W

F_B, the main effect of Factor B, = MS_B / MS_W

F_{AXB}, the interaction of Factor A and Factor B, = MS_{AXB} / MS_W

If there is no treatment effect for Factor A, then the ratio of MS_A / MS_W should be approximately 1.00. If there is a treatment effect, this F ratio will be greater than 1.00. The same will be true for the F ratio for Factor B and the F ratio for the interaction of Factor A and Factor B.

Conducting A Two-Way Between-Subjects ANOVA

For our first example of a two-way between-subjects ANOVA we will analyze a hypothetical set of data examining the effect of gender and age upon the likelihood of receiving traffic tickets. In our study, gender (Factor A) and age (Factor B) are the IVs. The DV is the number of traffic tickets received in the preceding three-year period. The null hypothesis for Factor A is that there is no difference between the number of tickets received by men and women. The null hypothesis for Factor B is that there is no difference between the number of tickets received by three different age groups of drivers: young, middle-aged and old. Finally, our null hypothesis for the interaction of Factor A and Factor B is that there is no unique effect of any combination of treatment levels. As usual we set our α equal to .05. As subjects are not being randomly assigned to treatment levels this is a quasi-experimental design.

Since our study consists of two levels of Factor A (men and women) and three levels of Factor B (young, middle-aged and old drivers), and each subject receives only one combination of treatment levels, this is a 2 X 3 between-subjects ANOVA. The data for the six combinations of gender and age, as well as the initial calculations for the ANOVA, are shown in Table 13.3. And note that Factor A has been assigned to the rows and Factor B to the columns in Table 13.3. Finally, to simplify the calculations the total number of subjects is unrealistically small.

Table 13.3 Example 2: Data and Initial Calculations

Factor A Factor B

	Young	Middle-Aged	Old	Row Totals

Men	X 9 6 6 5 $\Sigma X_{cell\ 1} = 26$ $n_{cell\ 1} = 4$ $M_{cell\ 1} = 6.50$	X 5 3 2 2 $\Sigma X_{cell\ 2} = 12$ $n_{cell\ 2} = 4$ $M_{cell\ 2} = 3.00$	X 6 5 5 4 $\Sigma X_{cell\ 3} = 20$ $n_{cell\ 3} = 4$ $M_{cell\ 3} = 5.00$	$\Sigma X_{row} = 58$ $n_{row} = 12$ $M_{row} = 4.83$
Women	X 9 7 5 4 $\Sigma X_{cell\ 4} = 25$ $n_{cell\ 4} = 4$ $M_{cell\ 4} = 6.25$	X 5 2 2 2 $\Sigma X_{cell\ 5} = 11$ $n_{cell\ 5} = 4$ $M_{cell\ 5} = 2.75$	X 7 6 4 4 $\Sigma X_{cell\ 6} = 21$ $n_{cell\ 6} = 4$ $M_{cell\ 6} = 5.25$	$\Sigma X_{row} = 57$ $n_{row} = 12$ $M_{row} = 4.75$
Column Totals	$\Sigma X_{col} = 51$ $n_{col} = 8$ $M_{col} = 6.38$	$\Sigma X_{col} = 23$ $n_{col} = 8$ $M_{col} = 2.88$	$\Sigma X_{col} = 41$ $n_{col} = 8$ $M_{col} = 5.13$	$\Sigma X_{total} = 115$ $n_{total} = 24$ $M_{total} = 4.79$

The next step is to create a table showing the seventeen values that must be found in the calculation of a two-way between-subjects ANOVA (Table 13.4).

Table 13.4 Example 2: Summary Table for the Two-way Between-subjects ANOVA

Source of Variation	SS	df	MS	F
Factor A	SS_A	df_A	MS_A	F ratio
Factor B	SS_B	df_B	MS_B	F ratio
Interaction AXB	SS_{AXB}	df_{AXB}	MS_{AXB}	F ratio
Within Groups	SS_W	df_W	MS_W	
Total	SS_T	df_T		

We then proceed essentially as with the one-way between-subjects ANOVA except that additional calculations are needed to determine Factor A, Factor B and the Interaction AXB. These calculations are not difficult. However, they are time-consuming. Anyone conducting a two-factor or two-way between-subjects ANOVA is strongly encouraged to utilize a statistical package such as SPSS. Accordingly, the outcomes rather than the actual steps of calculating the sums of squared deviations will be presented. Then the remaining steps to complete the ANOVA table will be described. As each value is calculated, it is entered into Table 13.5.

The Sums Of Squares

We begin by noting that we will need five values for SS, then we find five values for df, four values for MS, and finally three F ratios.

Recall that in a one-way between-subjects ANOVA:

$$SS_T = SS_{Bet} + SS_W$$

As was noted previously, with a two-way between-subjects ANOVA, the SS_{Bet} is partitioned, or divided, into three parts:

$$SS_{Bet} = SS_A + SS_B + SS_{AXB}$$

Thus, for a two-way between-subjects ANOVA we have:

$$SS_T = SS_A + SS_B + SS_{AXB} + SS_W$$

The SS for these data using SPSS are:

$SS_A = 0.04$

$SS_B = 50.33$

$SS_{AXB} = 0.33$

$SS_W = 45.25$

$SS_T = 95.96$

As a check on our calculations, we note that:

$SS_T = SS_A + SS_B + SS_{AXB} + SS_W$

$95.96 = 0.04 + 50.33 + 0.33 + 45.25$

$95.96 = 95.95$ except for minor rounding error when reducing to two decimal places

Calculating Degrees Of Freedom

We now must calculate the degrees of freedom for Factor A, Factor B, the Interaction AXB, Within Groups and Total:

df_A = Number of levels of Factor A (in our example, the number of rows) – 1

$= 2 – 1$

$= 1$

df_B = Number of levels of Factor B (in our example, the number of columns) – 1

$= 3 - 1$

$= 2$

$df_{AXB} = df_A \text{ X } df_B$

$= 1 \text{ X } 2$

$= 2$

df_W = N – the number of cells

where N is the total number of subjects in the study.

$= 24 - 6$

$= 18$

$df_T = N - 1$

$= 24 - 1$

$= 23$

As a check on our calculations:

$df_T = df_A + df_B + df_{AXB} + df_W$

$23 = 1 + 2 + 2 + 18$

$23 = 23$

Calculating Mean Squares

The MS for Factor A, Factor B, the Interaction AXB, and the MS_W are found by dividing the appropriate SS by its df. Thus:

$MS_A = \frac{\mathbf{SS_A}}{\mathbf{df_A}} = \frac{\mathbf{0.04}}{\mathbf{1}} = 0.04$

$MS_B = \frac{\mathbf{SS_B}}{\mathbf{df_B}} = \frac{\mathbf{50.33}}{\mathbf{2}} = 25.17$

$MS_{AXB} = \frac{\mathbf{SS_{AXB}}}{\mathbf{df_{AXB}}} = \frac{\mathbf{0.33}}{\mathbf{2}} = 0.17$

$MS_W = \frac{\mathbf{SS_W}}{\mathbf{df_W}} = \frac{\mathbf{45.25}}{\mathbf{18.00}} = 2.51$

Calculating The F Ratios

The F ratios for Factor A, Factor B, and the Interaction AXB are found by dividing each of their MS by the MS_W. Thus:

$F_A = \frac{\mathbf{MS_A}}{\mathbf{MS_W}} = \frac{\mathbf{0.04}}{\mathbf{2.51}} = .02$

$F_B = \frac{\mathbf{MS_B}}{\mathbf{MS_W}} = \frac{\mathbf{25.17}}{\mathbf{2.51}} = 10.03$

$F_{AXB} = \frac{\mathbf{MS_{AXB}}}{\mathbf{MS_W}} = \frac{\mathbf{0.17}}{\mathbf{2.51}} = .07$

With the calculation of these three F values Table 13.5 is complete. We will later calculate the values in the final columns of this table.

Table 13.5 Example 2: Completed Summary Table for the Two-way Between-subjects ANOVA, with the Values for Partial Eta Squared (η_p^2) and Eta Squared (η^2)

Source of Variation	SS	df	MS	F	η_p^2	η^2
Factor A	0.04	1	0.04	0.02		
Factor B	50.33	2	25.17	10.03**	0.53	0.52
AXB	0.33	2	0.17	0.07		
Within	45.25	18	2.51			
Total	95.96*	23				

*Single asterisk indicates there is minor rounding error.

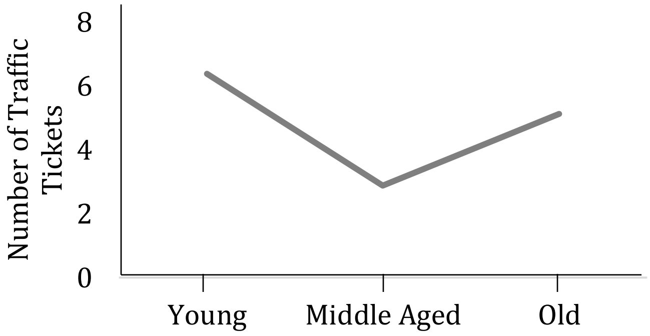
**Double asterisk indicates the F ratio is larger than the critical value for an α of .01.

Interpreting The F Ratio

We must enter the F table (Appendix K, Table 4) to determine whether any of these three F ratios is significantly different from a value of 1.00, which would be expected if the null hypotheses were true. Remember, each F ratio is based upon two MS estimates of the population variance. To find the critical value of F, we locate the column in the F table corresponding to the df associated with the MS of our numerator, and the row corresponding to the df associated with the MS of our denominator. For the F ratio of Factor A these are 1 and 18 df. At the intersection of our column and row in the F table we find the critical value with an α of .05 is 4.41. As the obtained value of F for Factor A is 0.02, this is not statistically significant. (Remember, any F ratio less than 1.0 will not be statistically significant.) For Factor B and the Interaction AXB, the df would be 2 and 18. The critical value for an α of .05 is 3.55. As the obtained value of F for Factor B is 10.03 it is statistically significant. (In fact, a more comprehensive table would indicate that the critical value for an α of .01 is 6.01. Thus, Factor B is also statistically significant at the .01 level. This is indicated by ** in Table 13.5.) Finally, the obtained value of F for the Interaction AXB is 0.07, which is not statistically significant.

The statistically significant main effect for Factor B can be presented visually (Figure 13.4) by graphing the three column means we calculated in Table 13.3.

Figure 13.4 Example 2: Graph of the Significant Main Effect of Age for the Number of Traffic Tickets



Conducting The Post Hoc Comparisons

There are three levels to Factor B. While our significant F indicates that there was an effect for age (this is a main effect), it does not specify which treatment levels of Factor B (column means) differ. In order to make this determination we need to conduct post hoc comparisons.

The number of main effect pairwise comparisons is given by the equation:

$$\text{Number of pairwise comparisons} = \frac{\mathbf{k(k - 1)}}{\mathbf{2}}$$

where k is the number of means being considered. With a 2 X 3 ANOVA, k (remember we are now comparing the column means) equals 3. Thus, there are [3(3 – 1)] / 2 = 3, pairwise comparisons. These 3 pairwise comparisons between the column means are:

Difference between the means of the young and middle-aged subjects = 6.38 – 2.88 = 3.50

Difference between the means of the young and old subjects = 6.38 – 5.13 = 1.25

Difference between the means of the middle-aged and old subjects = 2.88 – 5.13 = –2.25

(As before, we ignore the sign of the differences as the sign simply reflects the order the means were subtracted.)

The significant F for Factor B indicates that at least one of these three differences is expected to be statistically significant, in other words, not due to chance. To specify which means differ, we will once again use Tukey's HSD test.

For a significant *main effect* in a two-way between-subjects ANOVA, the critical value for the Tukey HSD test is found using the same equation as for a one-way between-subjects ANOVA:

$$\text{Critical value of Tukey HSD} = q\sqrt{\frac{\mathbf{MS_W}}{\mathbf{n}}}$$

Where n = the number of scores for *each* mean (which equals 8 as we are dealing with column totals). (This definition of n assumes that each mean is based upon an equal n.) The value of MS_W comes from the ANOVA table.

The value for q is found in the q table (Appendix K, Table 5). The column to use is determined by the number of levels of the IV (number of means being compared), in our case 3.

The row is determined by the degrees of freedom of the MS_W, in our case 18. With α equal to .05, q is equal to 3.61. We can now find the critical value for the Tukey HSD test:

$$\text{Critical value of Tukey HSD with } \alpha \text{ equal to .05} = 3.61\sqrt{\frac{\mathbf{2.51}}{\mathbf{8}}}$$

$$= 3.61\sqrt{0.31}$$

$$= (3.61)(0.56)$$

$$= 2.02$$

This indicates that the pairwise comparisons between column means must result in a difference *as great or greater* than this critical value of 2.02 in order to be considered significant. As is evident from our list of pairwise differences, the comparison between the young and middle-aged subjects, as well as the comparisons between the middle-aged and old subjects are statistically significant. However, with these hypothetical data the pairwise comparison between the young and old subjects is not statistically significant.

With α equal to .01, q is equal to 4.70, and the critical value then becomes:

$$\text{Critical value of Tukey HSD with } \alpha \text{ equal to .01} = 4.70\sqrt{\frac{\mathbf{2.51}}{\mathbf{8}}}$$

$$= 4.70\sqrt{0.31}$$

$$= (4.70)(0.56)$$

$$= 2.63$$

As the comparison between the young and middle-aged subjects exceeds this critical value of 2.63 it is also statistically significant at the .01 level.

Thus far in the analysis we have determined that only one of the F ratios (for Factor B) was statistically significant, and we have conducted our post hoc comparisons. We now turn to the determination of effect sizes. (Note that following a significant F ratio it does not matter whether you conduct the post hoc comparisons or find the effect sizes first.)

Calculating The Effect Size

SPSS calculates a partial eta squared ($\eta_{\mathbf{p}}^{\mathbf{2}}$) for each of the F ratios of a two-way between-subjects ANOVA. The equations for the three $\eta_{\mathbf{p}}^{\mathbf{2}}$ values can be written as:

$$\eta_{\mathbf{p}}^{\mathbf{2}} \text{ for Factor A} = \frac{\mathbf{SS_A}}{\mathbf{SS_T - SS_B - SS_{AXB}}}$$

$$\eta_{\mathbf{p}}^{\mathbf{2}} \text{ for Factor B} = \frac{\mathbf{SS_B}}{\mathbf{SS_T - SS_A - SS_{AXB}}}$$

$$\eta_{\mathbf{p}}^{\mathbf{2}} \text{ for Interaction AXB} = \frac{\mathbf{SS_{AXB}}}{\mathbf{SS_T - SS_A - SS_B}}$$

As only the main effect for factor B was significant, we would only report (or calculate by hand) a measure of effect size for this component of the ANOVA:

$$\eta_{\mathbf{p}}^{\mathbf{2}} \text{ for Factor B} = \frac{\mathbf{SS_B}}{\mathbf{SS_T - SS_A - SS_{AXB}}} = \frac{\mathbf{50.33}}{\mathbf{95.96 - 0.04 - 0.33}} = \frac{\mathbf{50.33}}{\mathbf{95.59}} = 0.53 = 53\%$$

The $\eta_{\mathbf{p}}^{\mathbf{2}}$ for the significant F ratio is included in Table 13.5. $\eta_{\mathbf{p}}^{\mathbf{2}}$ provides a measure of the proportion of total variability being accounted for after subtracting the variability associated with the other components of the ANOVA. A characteristic of $\eta_{\mathbf{p}}^{\mathbf{2}}$ is that the sum of $\eta_{\mathbf{p}}^{\mathbf{2}}$ for the different components of an ANOVA *may not* equal 1.00. Thus you cannot check your calculations by adding the values for each component of the ANOVA.

It has been suggested that eta squared (η^2) should be used instead of $\eta_{\mathbf{p}}^{\mathbf{2}}$, or that both measures of effect size should be reported (Levine & Hullett, 2002). With a two-way between-subjects ANOVA, an η^2 indicates the percent of variability explained by each of the main effects and the interaction. However, you must calculate η^2 by hand as SPSS does not provide η^2 for a two-way between-subjects ANOVA. The equations and calculations for the three η^2 values associated with the F ratios are:

$$\eta^2 \text{ for Factor A} = \frac{\mathbf{SS_A}}{\mathbf{SS_T}} = \frac{\mathbf{0.04}}{\mathbf{95.96}} = .00 = 0\%$$

$$\eta^2 \text{ for Factor B} = \frac{\mathbf{SS_B}}{\mathbf{SS_T}} = \frac{\mathbf{50.33}}{\mathbf{95.96}} = 0.52 = 52\%$$

$$\eta^2 \text{ for Interaction AXB} = \frac{\mathbf{SS_{AXB}}}{\mathbf{SS_T}} = \frac{\mathbf{0.33}}{\mathbf{95.96}} = .00 = 0\%$$

In addition, an η^2 for the within component of the ANOVA can also be calculated:

$$\eta^2 \text{ for within} = \frac{\mathbf{SS_W}}{\mathbf{SS_T}} = \frac{\mathbf{45.25}}{\mathbf{95.96}} = 0.47 = 47\%$$

As a check on our calculations, these four values of η^2 should sum to 1.00. This is confirmed below except for minor rounding error:

$$0.00 + 0.52 + 0.00 + 0.47 \approx 1.00$$

However, in our example only one main effect was statistically significant and thus we would only report an η^2 value for Factor B in a paper. This value of 0.52 is listed in the final column of Table 13.5.

It is important to note that while for our example $\eta_{\mathbf{p}}^{\mathbf{2}}$ and η^2 are virtually identical, the differences between $\eta_{\mathbf{p}}^{\mathbf{2}}$ and η^2 can be substantial. This is because while each is providing a measure of effect size they are providing different information. With η^2 we are finding the percent of the *total variability* explained by each factor, the interaction, and within. In contrast, with $\eta_{\mathbf{p}}^{\mathbf{2}}$ we are finding the percent of the *unaccounted for variability* that is explained by each factor and the interaction. Put another way, $\eta_{\mathbf{p}}^{\mathbf{2}}$ is measuring the percent of variance explained after removing the other sources of variability. Thus, when calculating the $\eta_{\mathbf{p}}^{\mathbf{2}}$ for Factor A, the variability accounted for by Factor B and by the interaction is removed from the denominator. The result is a measure of the remaining variability that is accounted for by Factor A. The same is occurring with the $\eta_{\mathbf{p}}^{\mathbf{2}}$ for Factor B and for the Interaction AXB.

Reporting The Results Of A Two-Way Between-Subjects ANOVA With A Significant Main Effect

In a paper, we would indicate the degrees of freedom used and the F ratios that were obtained. For each main effect that was significant we would also include the significant post hoc pairwise comparisons and indicate the measure of effect size. Specifically, based upon our calculations we would report that neither the main effect for gender (Factor A) nor the Interaction AXB was found to be significant ($F(1,18) = 0.02$, $p > .05$ and $F(2,18) = 0.07$, $p > .05$, respectively). (Note the direction of the $>$ symbol.) However the main effect for age was significant $F(2,18) = 10.03$, $p < .01$, $\eta_p^2 = .53$). (Alternatively, we could report η^2, or both η^2 and η_p^2.) Finally, we would report that the Tukey's HSD test was conducted and we would indicate that, overall, the young and old drivers received more tickets than the drivers who were middle-aged.

At the end of this chapter SPSS is used to analyze these data. The outcome is almost identical, though with SPSS there is greater precision and exact p-values are given. Specifically, based upon the SPSS analysis we would once again indicate that neither the main effect for Factor A (gender) nor the Interaction AXB was found to be significant ($F(1,18) = 0.02$, $p = .899$ and $F(2,18) = 0.07$, $p = .936$, respectively). However the main effect for Factor B (age) was significant $F(2,18) = 10.01$, $p = .001$, $\eta_p^2 = .527$). Note that our decision to reject the null hypothesis is confirmed for Factor B as the p-value was less than our α of .05. In addition, SPSS calculates both Levene's test and Tukey's HSD test.

Progress Check

1. A two-way between subjects ANOVA with 4 levels of one IV and 6 for the other IV would be called a _____ ANOVA.
2. With a two-way between-subjects ANOVA there are _____ main effects, and a total of _____ F ratios are calculated.
3. If an outcome is due to a particular combination of the independent variables, this is an example of a(an) _____.

Answers: 1. 4 X 6 2. two; three 3. interaction

A Second Example

For our second example of a two-way between-subjects ANOVA we will analyze data from a hypothetical experiment on background music and studying. In our study, the IVs are the subject's history of living in a quiet or loud environment (Factor A) and the presence or absence of background music while studying (Factor B). The DV is the subsequent quiz grade. The null hypothesis for Factor A is that there is no difference in quiz grades between subjects who lived in quiet environments versus those who lived in loud environments. The null hypothesis for Factor B is that there is no difference in quiz grades as a result of studying with or without background music. Finally, our null hypothesis for the interaction of Factor A and Factor B is that there is no unique effect of any combination of treatment levels. As usual we set our α equal to .05 for each of the three hypotheses we will be testing.

Sharp-eyed readers will have noted that while subjects can be randomly assigned to study with or without background music (Factor B), they cannot be randomly assigned to a history of exposure to sound (Factor A). Thus Factor A is quasi-experimental while Factor B is a true experimental variable. This does not affect the statistical analysis, but can affect the researcher's interpretation of a significant outcome.

Since our study consists of two levels of Factor A (quiet or loud) and two levels of Factor B (music or no music), this is a 2 X 2 ANOVA and there are four combinations of the two IVs. These four combinations, along with the initial calculations that will be needed, are shown in Table 13.6. Thus, with a 2 X 2 ANOVA there are four cells. The two-way between-subjects ANOVA requires that there be an approximately equal number of data points in each cell. And, it is important to note that Factor A (quiet or loud history) was assigned to be the rows and Factor B (level of background music) is the columns in Table 13.6. Which factor is assigned to be rows and which is columns is arbitrary, but during the calculations it is critical to remember how you assigned your variables. Finally, note that the total number of subjects is only 12. This unrealistically small number was chosen in order to aid in showing the calculations.

Table 13.6 Example 2: Data and Initial Calculations

	Factor B		
Factor A			
	Music	No Music	Row Totals
Quiet History	X 4 4 3 $\Sigma X_{\text{cell 1}} = 11$ $n_{\text{cell 1}} = 3$ $M_{\text{cell 1}} = 3.67$	X 8 7 7 $\Sigma X_{\text{cell 2}} = 22$ $n_{\text{cell 2}} = 3$ $M_{\text{cell 2}} = 7.33$	$\Sigma X_{\text{row}} = 33$ $n_{\text{row}} = 6$ $M_{\text{row}} = 5.5$

Loud History	X 8 7 6 $\Sigma X_{\text{cell 3}} = 21$ $n_{\text{cell 3}} = 3$ $M_{\text{cell 3}} = 7.00$	X 4 3 2 $\Sigma X_{\text{cell 4}} = 9$ $n_{\text{cell 4}} = 3$ $M_{\text{cell 4}} = 3.00$	$\Sigma X_{\text{row}} = 30$ $n_{\text{row}} = 6$ $M_{\text{row}} = 5.00$
Column Totals	$\Sigma X_{\text{col}} = 32$ $n_{\text{col}} = 6$ $M_{\text{col}} = 5.33$	$\Sigma X_{\text{col}} = 31$ $n_{\text{col}} = 6$ $M_{\text{col}} = 5.17$	$\Sigma X_{\text{total}} = 63$ $n_{\text{total}} = 12$ $M_{\text{total}} = 5.25$

The next step is to create a table showing what it is that must be calculated. It is the same as for our previous example (Table 13.4).

Recall that in a two-way between-subjects ANOVA the SS_{Bet} is partitioned into SS_A, SS_B and SS_{AXB}. Therefore, to complete the table we need to have these three values for SS, as well as the SS_W and the SS_T, then find the five values for df, the four values for MS and finally three F ratios. As each value is determined it is entered into the ANOVA summary table (Table 13.7). And as with the previous example we will not actually calculate the values for SS as this is tedious when done by hand.

The Sums Of Squares

For our data, the values for the five needed SS, calculated with SPSS, are:

$SS_A = 0.75$

$SS_B = 0.08$

$SS_{AXB} = 44.08$

$SS_W = 5.33$

$SS_T = 50.25$

As a check, we note that with a two-way between-subjects ANOVA:

$SS_T = SS_A + SS_B + SS_{AXB} + SS_W$

$50.25 = 0.75 + 0.08 + 44.08 + 5.33$

$50.25 = 50.24$ except for minor error due to rounding

Calculating Degrees Of Freedom

We now must calculate the values for the degrees of freedom that correspond to each SS:

df_A = Number of levels of Factor A (in our example, the number of rows) – 1

$= 2 - 1$

$= 1$

df_B = Number of levels of Factor B (in our example, the number of columns) – 1

$= 2 - 1$

$= 1$

$df_{AXB} = df_A \text{ X } df_B$

$= 1 \text{ X } 1$

$= 1$

df_W = N – the number of cells

where N is the total number of subjects in the study.

$= 12 - 4$

$= 8$

$df_T = N - 1$

$= 12 - 1$

$= 11$

As a check on our calculations,

$df_T = df_A + df_B + df_{AXB} + df_W$

$11 = 1 + 1 + 1 + 8$

$11 = 11$

Calculating Mean Squares

The MS for Factor A, Factor B, the interaction and the MS_W are found by dividing the appropriate SS by its degrees of freedom. Thus:

$MS_A = \frac{\mathbf{SS_A}}{\mathbf{df_A}}$ $MS_B = \frac{\mathbf{SS_B}}{\mathbf{df_B}}$ $MS_{AXB} = \frac{\mathbf{SS_{AXB}}}{\mathbf{df_{AXB}}}$ $MS_W = \frac{\mathbf{SS_W}}{\mathbf{df_W}}$

$= \frac{\mathbf{0.75}}{\mathbf{1}}$ $= \frac{\mathbf{0.08}}{\mathbf{1}}$ $= \frac{\mathbf{44.08}}{\mathbf{1}}$ $= \frac{\mathbf{5.33}}{\mathbf{8}}$

$= 0.75$ $= 0.08$ $= 44.08$ $= 0.67$

Calculating The F Ratios

The F ratios for Factor A, Factor B, and the Interaction are found by dividing each MS by MS_W. Thus:

$$F_A = \frac{\mathbf{MS_A}}{\mathbf{MS_W}} \qquad F_B = \frac{\mathbf{MS_B}}{\mathbf{MS_W}} \qquad F_{AXB} = \frac{\mathbf{MS_{AXB}}}{\mathbf{MS_W}}$$

$$= \frac{\mathbf{0.75}}{\mathbf{0.67}} \qquad = \frac{\mathbf{0.08}}{\mathbf{0.67}} \qquad = \frac{\mathbf{44.08}}{\mathbf{0.67}}$$

$$= 1.12 \qquad = 0.12 \qquad = 65.79$$

With the calculation of these three F values our summary table is complete (Table 13.7). We will later calculate the values in the final two columns of this table.

Table 13.7 Example 1: Completed Summary Table for the Two-way Between-subjects ANOVA, with the Values for Partial Eta Squared (η_p^2) and Eta Squared (η^2)

Source of Variation	SS	df	MS	F	η_p^2	η^2
Factor A	0.75	1	0.75	1.12		
Factor B	0.08	1	0.08	0.12		
AXB	44.08	1	44.08	65.79*	0.89	0.88
Within	5.33	8	0.67			
Total	50.25	11				

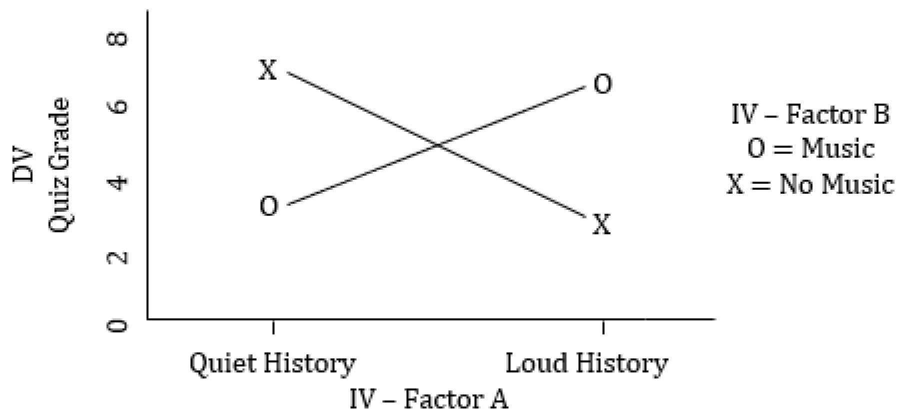
*Asterisk indicates the F ratio for the interaction is larger than the critical value for an α of .05.

Interpreting The F Ratios

To determine whether any of these three F ratios is significantly different from the expected value of 1.00 we must enter the F table. Remember, the F ratio is based upon two values of MS, each with its degrees of freedom. To find the critical value of F we locate the column in the F table corresponding to the degrees of freedom in the numerator of our F ratio and the row corresponding to the degrees of freedom in the denominator of the F ratio. Because this is a 2 X 2 ANOVA, all three of our F ratios are based on the same numbers of degrees of freedom. In this example the degrees of freedom are 1 and 8. As usual, we have chosen an α of .05. At the intersection of our column and row in the F table (Appendix K, Table 4) we find the critical value of 5.32. Only the F ratio for the Interaction AXB is larger than this critical value. This is indicated with an * in Table 13.7. We therefore conclude that there were no significant main effects but there was a significant interaction.

This significant interaction is presented in Figure 13.5. Each point in this figure is a cell mean from Table 13.6.

Figure 13.5 Example 2: Graph of the Interaction



While our significant F indicates that there was an interaction, it does not specify which cell means differ, or the effect size for the interaction.

Conducting The Post Hoc Comparisons

Post hoc comparisons are generally only conducted following a significant F ratio. In our current example neither main effect was statistically significant and thus no post hoc comparison would be conducted for them. However, even if one or both of our main effects was significant, post hoc comparisons would still not be conducted. This is because each IV, history of living in a quiet or loud environment, and level of background music while studying, had only two levels. Thus there would be no need to calculate a post hoc test for these comparisons - inspection of the data would indicate the nature of any observed difference. However, the interaction was significant and it involves four cell means. Thus, post hoc comparisons will be needed in order to specify where the effect is. In addition, with a factorial ANOVA, if the interaction is statistically significant then the focus is upon the interaction even if one or both main effects is statistically significant.

As you may recall, the number of comparisons between means, called pairwise comparisons, in an experiment is given by the equation:

$$\text{Number of pairwise comparisons} = \frac{\mathbf{k(k-1)}}{\mathbf{2}}$$

where k is the number of means being compared.

In our case, k equals 4 as we are interested in the four cell means since we are dealing with a significant interaction. These cell means are presented, along with the number of each cell, in Table 13.8. There are [4(4 – 1)] / 2, which equals 6, pairwise comparisons. The 6 pairwise comparisons between the four cell means are shown in Table 13.9. (Determination of which of these comparisons is statistically significant will be described shortly. And remember, when comparing the differences between these means we ignore the sign of the difference as this simply indicates the order of the subtraction of the means.)

Table 13.8 Example 2: Cell Means

	Music	No Music
Quiet History	Cell 1 M = 3.67	Cell 2 M = 7.33
Loud History	Cell 3 M = 7.00	Cell 4 M = 3.00

Table 13.9 Example 2: Differences Between Cell Means – Significant Post Hoc Comparisons are Noted (The Two Confounded Comparisons are Italicized)

Mean cell 1 – mean cell 2	3.67 – 7.33 = -3.66**
Mean cell 1 – mean cell 3	3.67 – 7.00 = -3.33**
Mean cell 1 – mean cell 4	*3.67 – 3.00 = 0.67*
Mean cell 2 – mean cell 3	*7.33 – 7.00 = 0.33*
Mean cell 2 – mean cell 4	7.33 – 3.00 = 4.33**
Mean cell 3 – mean cell 4	7.00 – 3.00 = 4.00**

**Double asterisk indicates that the difference between cell means is larger than the critical value for an α of .01.

However, it is important to note that we cannot interpret all 6 of these comparisons of cell means. We can only interpret those comparisons in which just one of the IVs is varying. For instance, in the comparison between the mean of cell 1 and the mean of cell 2 (Table 13.8), the difference between the mean of 3.67 (mean of the group with a quiet history who heard music in the background during the experiment) and the mean of 7.33 (mean of the group with a quiet history who did not hear music in the background) refers to the effect of hearing different levels of background music upon subjects who all had a quiet history. Thus only one IV (level of background music during the experiment) is varying and the comparison can be interpreted. Similarly, the mean of cell 1, which is 3.67, can be meaningfully compared with the mean of cell 3, which is 7.00, since the groups differed on their history of sound exposure, but both groups heard the same level of background music during the experiment. However, the mean of cell 1, which is 3.67, (mean of the group with a quiet history who heard music in the background during the experiment) cannot be meaningfully compared with the mean of cell 4, which is 3.00, (mean of the group with a loud history who did not hear music in the background during the experiment) as in this case both IVs (history and current exposure) are varying.

Comparisons of cell means in which only one IV (factor) is varying are called **unconfounded comparisons**. These comparisons can be interpreted. If a comparison of cell means involves two IVs (factors) that are changing, this is called a **confounded comparison** and the outcome cannot be interpreted. Put another way, when referring to Tables 13.6 and 13.8 any difference between cell means that involves a vertical or horizontal comparison is unconfounded and can be interpreted. Any difference between cell means that involves a diagonal comparison is confounded and cannot be interpreted. (This is explained further in the box below.) The two confounded comparisons are indicated in Table 13.9 by being italicized.

Unconfounded comparison – Comparison of two cell means which involves only one factor that is changing. The comparison can be interpreted.

Confounded comparison – Comparison of two cell means which involves two factors that are changing. The comparison cannot be interpreted.

Of course, we still don't know which of these cell means differ significantly. The significant F simply indicates that we expect at least one of the cell means differs from another. To specify which cell means differ we need to conduct a post hoc test. Fortunately, we can again use Tukey's HSD test, though it will need to be modified slightly when dealing with a significant interaction.

As you will recall, calculation of the Tukey HSD leads to a critical value that is compared to the difference between each pair of means. Specifically, for a significant *interaction*:

$$\text{Critical value of Tukey HSD} = q_i\surd\frac{\mathbf{MS_W}}{\mathbf{n}}$$

where q_i is based upon the number of unconfounded comparisons of cell means in the interaction and is derived from q (refer to a more advanced statistical text for further details on how to obtain the value of q_i), MS_W comes from the ANOVA table, and n equals the number of scores for *each* cell mean (It is important to note that this equation for the critical value of the Tukey HSD is only appropriate for designs with an equal n for each cell.)

As there are four unconfounded comparisons of cell means in this example (Table 13.9) and there are 8 df for the MS_W, the value for q_i is 4.04. (Refer to a more advanced statistical text for further details on how to obtain this value.) We can now find the critical value for an α equal to .05:

$$\begin{aligned}\text{Critical value of Tukey HSD} &= 4.04\surd\frac{\mathbf{0.67}}{\mathbf{3}}\\ &= 4.04\surd 0.22\\ &= (4.04)(0.47)\\ &= 1.90\end{aligned}$$

Thus, in order to be considered significant with α equal to .05, the Tukey HSD test indicates that the difference between *unconfounded* cell means must be *as great or greater* than the critical value of 1.90.

We can also find the critical value for an α equal to .01. In this case the value for q_i is 5.63:

$$\text{Critical value} = 5.63\sqrt{\frac{\mathbf{0.67}}{\mathbf{3}}}$$

$$= 5.63\sqrt{0.22}$$

$$= (5.63)(0.47)$$

$$= 2.65$$

As all of the unconfounded comparisons of cell means have differences greater than 2.65 (Table 13.9), all of the pairwise comparisons of unconfounded cell means are significant at the .05 and at the .01 levels. An ** indicates those comparisons significant at the .01 level (Table 13.9). If any comparisons were significant at the .05 level but not at the .01 level, we might differentiate them by using an *.

Finally, it should be noted that SPSS does not calculate post hoc comparisons for an interaction. Instead, you can conduct your post hoc comparisons by hand as well as plot your cell means to assist in interpreting a significant interaction.

Box Dealing With Identifying Unconfounded And Confounded Comparisons

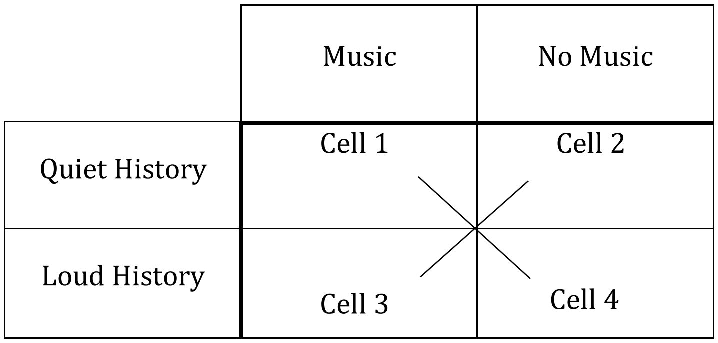
With the previous example of a 2 X 2 ANOVA it was noted that there were a total of six post hoc pairwise comparisons of cell means. However, only four of these comparisons could be interpreted as these only had one IV (factor) that was varying. These four comparisons are said to be unconfounded. And, as Table 13.10 indicates, each unconfounded comparison consists of either a vertical or horizontal comparison of cell means. These unconfounded comparisons are also listed in Table 13.9.

Table 13.10 Unconfounded Comparisons of the Cell Means of a 2 X 2 ANOVA

	Music	No Music
Quiet History	Cell 1	Cell 2
Loud History	Cell 3	Cell 4

It was also discussed that with a 2 X 2 ANOVA, two of the six post hoc pairwise comparisons of cell means involve both IVs (factors) varying. These comparisons cannot be interpreted and are said to be confounded. Table 13.11 indicates that each of these involves a diagonal comparison.

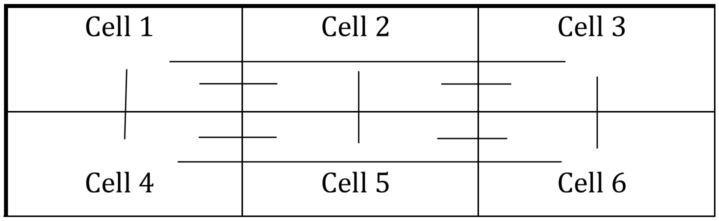
Table 13.11 Confounded Comparisons of the Cell Means of a 2 X 2 ANOVA



	Music	No Music
Quiet History	Cell 1	Cell 2
Loud History	Cell 3	Cell 4

The logic remains the same for larger ANOVAs. For a 2 X 3 ANOVA there are a total of 15 post-hoc pairwise comparisons of cell means. Of these, nine are unconfounded, and can be interpreted. For each of these comparisons only one IV (factor) is varying. As Table 13.12 indicates, each of these consists of either a vertical or horizontal comparison of cell means.

Table 13.12 Unconfounded Comparisons of the Cell Means of a 2 X 3 ANOVA



Cell 1	Cell 2	Cell 3
Cell 4	Cell 5	Cell 6

The nine unconfounded comparisons are identified in Table 13.13.

Table 13.13 The Nine Unconfounded Comparisons of Cell Means of a 2 X 3 ANOVA

Cell 1 versus Cell 2

Cell 2 versus Cell 3

Cell 4 versus Cell 5

Cell 5 versus Cell 6

Cell 1 versus Cell 4

Cell 2 versus Cell 5

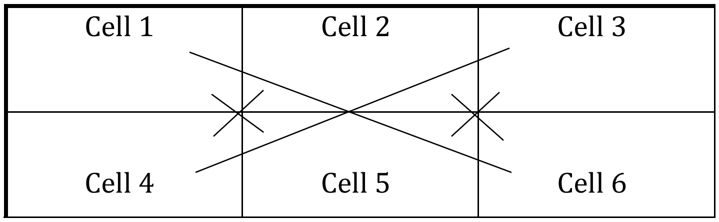
Cell 3 versus Cell 6

Cell 1 versus Cell 3

Cell 4 versus Cell 6

For a 2 X 3 ANOVA, there would be six post hoc pairwise comparisons of cell means that involve both IVs (factors) varying. These are called confounded comparisons and they cannot be interpreted. As Table 13.14 indicates, each of these consists of diagonal comparisons of cell means.

Table 13.14 Confounded Comparisons of the Cell Means of a 2 X 3 ANOVA



Cell 1	Cell 2	Cell 3
Cell 4	Cell 5	Cell 6

The six confounded comparisons are identified in Table 13.15.

Table 13.15 The Six Confounded Comparisons of Cell Means of a 2 X 3 ANOVA

Cell 1 versus Cell 5	Cell 3 versus Cell 5
Cell 2 versus Cell 4	Cell 1 versus Cell 6
Cell 2 versus Cell 6	Cell 3 versus Cell 4

Calculating The Effect Size

To this point in the analysis we have found that only the F ratio for the interaction was statistically significant and we have conducted our post hoc comparisons of cell means. We now need to calculate a measure of the effect size. (Note that it is arbitrary whether you begin by conducting the post hoc comparisons or find the effect size first following the determination that you have a significant F ratio.)

As was noted previously, SPSS utilizes **partial eta squared** ($\eta_{\mathbf{p}}^{\mathbf{2}}$) as a measure of effect size for a two-way between-subjects ANOVA. For a two-way between-subjects ANOVA the equations for $\eta_{\mathbf{p}}^{\mathbf{2}}$ can be written as:

$$\eta_{\mathbf{p}}^{\mathbf{2}} \text{ for Factor A} = \frac{\mathbf{SS_A}}{\mathbf{SS_T - SS_B - SS_{AXB}}}$$

$$\eta_{\mathbf{p}}^{\mathbf{2}} \text{ for Factor B} = \frac{\mathbf{SS_B}}{\mathbf{SS_T - SS_A - SS_{AXB}}}$$

$$\eta_{\mathbf{p}}^{\mathbf{2}} \text{ for Interaction AXB} = \frac{\mathbf{SS_{AXB}}}{\mathbf{SS_T - SS_A - SS_B}}$$

In our example, as only the interaction was found to be significant we would only report (or calculate by hand) one $\eta_{\mathbf{p}}^{\mathbf{2}}$:

$$\eta^2_p \text{ for Interaction} = \frac{44.07}{50.25 - 0.75 - 0.09} = \frac{44.07}{49.41} = 0.89 \text{ or } 89\%$$

which is a very large value for η^2_p. This value is included in Table 13.7.

η^2_p provides a measure of the proportion of total variability accounted for after subtracting the variability associated with other components of the ANOVA. And as was noted previously, the sum of η^2_p for the different components of an ANOVA *may not* equal 1.00.

It has been suggested that eta squared (η^2) should be used instead of η^2_p, or that both measures of effect size should be reported (Levine & Hullett, 2002). With a two-way between-subjects ANOVA, the values of η^2 indicate the percent of variability explained by each of the main effects and the interaction. And though SPSS does not provide η^2 for a two-way between-subjects ANOVA, these values are easy to calculate. The equations and calculations for each η^2 associated with an F ratio are:

$$\eta^2 \text{ for Factor A} = \frac{SS_A}{SS_T} = \frac{0.75}{50.25} = 0.01 \text{ or } 1\%$$

$$\eta^2 \text{ for Factor B} = \frac{SS_B}{SS_T} = \frac{0.08}{50.25} = 0.00 \text{ or } 0\%$$

$$\eta^2 \text{ for Interaction AXB} = \frac{SS_{AXB}}{SS_T} = \frac{44.08}{50.25} = 0\ .88 \text{ or } 88\%$$ (This is a very large value for η^2.)

In addition, an η^2 for the within component of the ANOVA could also be calculated:

$$\eta^2 \text{ for Within} = \frac{SS_W}{SS_T} = \frac{5.33}{50.25} = 0.11 = 11\%$$

As a check on our calculations, these four values of η^2 should sum to 1.00:

$$0.01 + 0.00 + 0.88 + 0.11 = 1.00$$

However, remember that in our example only the interaction was statistically significant and thus we would only include this value in the final column of Table 13.7. And only one η^2 value, for the interaction, would be reported in a paper.

Finally, it is important to recognize that while the numerators of the equations for η^2 and η^2_p **are the same**, the values of η^2 and η^2_p may differ dramatically since their denominators differ. With η^2 the denominator is always SS_T. With a factorial ANOVA, the denominator of η^2_p is not constant.

Reporting The Results Of A Two-Way Between-Subjects ANOVA With A Significant Interaction

In a paper, we would indicate the degrees of freedom used, the F ratios that were obtained as well as which F ratio was significant, that the Tukey HSD post hoc was used to determine which pairwise unconfounded comparisons of cell means were significantly different and the measure of effect size. Specifically we would report that the main effects for history of exposure, and whether subjects listened to music or not during the experiment, were not significant ($F(1,8) = 1.12$, $p > .05$ and $F(1,8) = .13$, $p > .05$), respectively. However, since the interaction was found to be significant, a measure of effect size for the interaction should then be reported ($F(1,8) = 65.78$, $p < .05$, $\eta_p^2 = .89$). (You could, instead, report η^2, or both η^2 and η_p^2.) We would then state that Tukey's HSD test indicated that all four of the unconfounded cell mean comparisons were statistically different.

If you use SPSS to calculate the F ratios you will find minor discrepancies due to rounding error in our calculations. Specifically, as before we would report that the main effects for history of exposure, and whether subjects listened to music or not, were not significant ($F(1,8) = 1.13$, $p = .320$ and $F(1,8) = .13$, $p = .733$), respectively. Note that in each case the p-value is greater than our α of .05. However, the interaction is still found to be significant ($F(1,8) = 66.13$, $p < .001$, $\eta_p^2 = .89$). This is indicated by our p-value being less than our α of .05. (Remember, instead of reporting η_p^2 we could report η^2, or both η^2 and η_p^2.)

A discussion of these hypothetical results would emphasize that whether background music hinders or enhances studying depends upon the subject's history of exposure to sound. Specifically, these hypothetical data would indicate that subjects with a history of living in a quiet environment find background music disruptive to studying whereas subjects with a history of living in an environment with more background sound find a quiet situation disruptive to studying.

Purpose And Limitations Of Using The Two-way Between-subjects ANOVA

1. *Test for difference.* The null hypotheses are that neither treatment has an effect, and there is no interaction. Therefore, if the null hypotheses are correct, any differences between the rows or column means, or between the cell means, are due to chance. The alternative hypotheses are that the treatments do have an effect and/or that they interact. Thus the two-way between-subjects ANOVA tests whether there are main effects as well as whether there is an interaction between the IVs.
2. *Does not provide a measure of effect size.* The two-way between-subjects ANOVA, like the one-way between-subjects ANOVA, is a test of significance. It indicates whether an outcome is likely to have occurred by chance. If an F ratio is significant a measure of effect size, such as eta squared (η^2) or partial eta squared (η_p^2), should be calculated.
3. *Compares two or more sample means for each main effect*. Each factor of the two-way between-subjects ANOVA must have at least two levels or there is no variable. However, there can theoretically be any number of levels greater than one. Of course, a study with a large number of levels for one or both of the factors would be unwieldy to conduct, though the ANOVA would handle the data without difficulty.

4. *Does not indicate where the difference is.* If an IV has more than two levels, then a significant main effect should be followed up with a post hoc procedure such as the Tukey HSD test. A significant interaction should also be followed up with a post hoc procedure such as the Tukey HSD test, but only the unconfounded comparisons can be interpreted. *Finally, it should be noted that SPSS does not calculate post hoc comparisons for an interaction. Instead, you can conduct your post hoc comparisons by hand as well as plot your cell means to assist in interpreting a significant interaction.*

Assumptions Of The Two-way Between-subjects ANOVA

1. *Interval or ratio data.* The data are on an interval or ratio scale of measurement.
2. *Random samples.* Each sample is drawn at random from a population.
3. *Normally distributed populations.* Each population from which a sample is drawn has a normal distribution of scores. However, as stated in the Central Limit Theorem, the ANOVA will be accurate so long as each sample size is at least 30. If the sample size is less than 30, then it is important that the underlying population be normally distributed.
4. *Population variances are equal.* The populations from which samples are drawn have equal variances.
5. *Each cell has an approximately equal number of subjects*

Conclusion

The two-way between-subjects ANOVA is a very flexible test. As you recall, the major advantage of the ANOVA is that it controls the experimentwise error rate while simultaneously comparing two or more sample means. The specific advantage of conducting a two-way ANOVA rather than two, one-way ANOVAs is that with one analysis you test two IVs instead of just one and, in addition, you test whether these IVs interact. Thus, the two-way ANOVA provides substantially more information than a one-way ANOVA, which is why it is such a popular statistical procedure. As you would expect, the assumptions of the two-way between-subjects ANOVA are very similar to those of the one-way between-subjects ANOVA.

Final Thoughts On The Relationship Between The One-Way Between-Subjects ANOVA, The One-Way Within-

Subjects ANOVA And The Two-Way Between-Subjects ANOVA

We have just finished our introduction to the ANOVAs. The three ANOVAs that we reviewed are among the most commonly used statistical procedures. Each shares characteristics with the others, which is a major advantage when first trying to master them. Of course, these same similarities can lead to challenges when trying to keep each type of ANOVA distinct. A comparison of the one-way between-subjects ANOVA, the one-way within-subjects ANOVA and the two-way between-subjects ANOVA is provided in Appendix M.

The one-way between-subjects ANOVA provides a foundation upon which the others build. This ANOVA is located in the middle of Appendix M, and it is bolded. The sources of variability for this ANOVA as well as the equation for its F ratio are provided in the top portion of the table. On the left side of the top portion of the table the same information is provided for the one-way within-subjects ANOVA. A number of differences should be noted in these two ANOVAs. First, with the between-subjects ANOVA we calculate the Between Groups variability. However, for a within-subjects ANOVA this variability is now labeled Between Treatments since the same, or related, subjects are used throughout a within-subjects study. In addition, the Within Groups variability of the between-subjects ANOVA is partitioned into pre-existing subject differences and residual error when we have a within-subjects ANOVA. It should also be noted that the pre-existing subject differences have been crossed out, indicating that this variability is removed from the analysis of a within-subjects ANOVA. Finally, the F ratio of the one-way within-subjects ANOVA reflects this reduction in variability by substituting MS_{Res} for MS_W.

The two-way between-subjects ANOVA, which is located on the right side of Appendix M, is also closely related to the one-way between-subjects ANOVA. In this case, the Between Groups variability of the one-way between-subjects ANOVA is partitioned into three components: Factor A, Factor B and the Interaction AXB. However, the Within Groups variability is not partitioned and thus the denominator of each of the three F ratios in a two-way between-subjects ANOVA remains MS_W.

As Appendix M indicates, if no F ratio is found to be statistically significant your analysis is complete. However, if the F ratio in either the one-way between-subjects ANOVA or the one-way within-subjects ANOVA is statistically significant, or if at least one F ratio in the two-way between-subjects ANOVA is significant then you need to proceed to the bottom portion of Appendix M. As was the case previously, the middle section is bolded and refers to the one-way between-subjects ANOVA, the left portion refers to the one-way within-subjects ANOVA, and the right portion to the two-way between-subjects ANOVA. In order to determine where the significant effect is located in a one-way between-subjects ANOVA, Appendix M indicates that you use the Tukey HSD test, and to do so you need to find the value of q from the q table. To find the effect size you conduct eta squared (η^2). For a one-way within-subjects ANOVA, to find where the significant difference is located you would conduct a series of dependent t tests and use the Bonferroni method to control the Type I error rate. And you utilize partial eta squared (η^2_p) instead of η^2 as your measure of effect size. For the two-way between-subjects ANOVA you once again use the Tukey HSD test when finding where the significant effect is located. Remember, however, that you need to use q_i following a significant interaction. Finally, for each significant F ratio in a two-way between-subjects ANOVA you have a choice between reporting η^2 or η^2_p or both.

Glossary Of Terms

Cell – A particular combination of treatment levels in a Factorial ANOVA.

Confounded comparison – Comparison of two cell means which involves two factors that are changing. The comparison cannot be interpreted.

Factorial ANOVA – An ANOVA with more than one independent variable.

Interaction – A change in the dependent variable that is due to the presence of a particular combination of independent variables.

Main effect – With a factorial ANOVA, another term for an independent variable or factor.

Two-way between-subjects ANOVA – An inferential procedure for comparing means from independent samples when there are two independent variables.

Unconfounded comparison – Comparison of two cell means which involves only one factor that is changing. The comparison can be interpreted.

References

Caspi, A., McClay, J., Moffitt, T., Mill, J., Martin, J., Craig, I. W., Taylor, A., Poulton, R., & Moffitt, T. (2002). Role of genotype in the cycle of violence in maltreated children. *Science, 297,* 851 – 854.

Levine, T. R. & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research, 28,* 612-625.

Questions – Chapter 13

(Answers are provided in Appendix J.)

1. With a 2 X 3 X 4 ANOVA there are _____ independent variables.
   a. One

b. Two
c. Three
d. Four

2. In a 2 X 3 ANOVA, the number 3 indicates that there are _____.
a. Three independent variables
b. Three levels to an independent variable
c. Three subjects in a condition
d. None of the above

3. An interaction occurs when _____.
a. A specific combination of factors determines the value of the dependent variable
b. A specific combination of factors influences the independent variable
c. More than one experimenter is collecting the data
d. A particularly important finding occurs

4. A main effect occurs when _____.
a. An interaction occurs
b. The dependent variable has an effect
c. An independent variable has an effect
d. All of the above

5. In a two-way between-subjects ANOVA we calculate F ratios for _____.
a. One Main effect and two interactions
b. Two main effects and two interactions
c. Two main effects and one interaction
d. None of the above

6. In a two-way between-subjects ANOVA we calculate _____ F ratios.
a. One
b. Two
c. Three
d. Four

7. In a two-way between-subjects ANOVA, each combination of treatment levels is a _____.
a. Cell
b. Level
c. Condition
d. Factor

8. Following a significant main effect or interaction with a 3 X 6 between-subjects ANOVA, you would consider _____.
a. Redoing the study with larger sample sizes
b. Increasing the number of independent variables in the study
c. Calculating Tukey's HSD
d. None of the above

For questions 9 and 10 assume that we examine the effect on exam scores of amount of studying (students are randomly assigned to study a little or a great deal) and class in college (freshman, sophomore, junior and senior).

9. If there was a significant interaction and we compared the cell means for studying a little among freshmen, sophomores, juniors and seniors, these would be _____.

a. confounded comparisons
b. unconfounded comparisons

10. If there was a significant interaction and we compared the cell means for studying a little among freshmen with studying a great deal among sophomores, this would be a (an) _____.
a. confounded comparison
b. unconfounded comparison

For questions 11 – 13 assume the we have conducted a 2 X 5 ANOVA.

11. How many cells means would there be?
a. 8
b. 10
c. 20
d. 45

12. If the main effect for the factor with 5 treatment levels is found to be statistically significant, but the interaction is not significant, how many post hoc pairwise comparisons would there be?
a. 8
b. 10
c. 20
d. 45

13. If the interaction but neither main effect is found to be statistically significant, how many post hoc pairwise comparisons would there be?
a. 8
b. 10
c. 20
d. 45

The data for problems 14 – 16 are similar to those used in Chapter 11 (questions 15 – 19) except that we now assume that the original 18 students were 9 males and 9 females. Assume that $SS_{Gender} = 50$, the $SS_{Background} = 12$, $SS_{Gender\ X\ Background} = 4$, and $SS_T = 96$. Compare your answers to what you calculated in Chapter 11.

		Level of Background Noise		
		Quiet	Moderate	Noisy
		9	7	6
	Women	10	9	8
		8	8	10
Gender				
		13	13	7
	Men	12	11	11
		14	12	12

14. What is the value of F for gender?
a. 50
b. 20
c. 0.8
d. 2.4

15. What is the value of F for background noise level?
   a. 50
   b. 20
   c. 0.8
   d. 2.4

16. What is the value of F for the interaction of gender X background?
   a. 50
   b. 20
   c. 0.8
   d. 2.4

Problems 17 – 24 utilize SPSS.

Using SPSS With The Two-way Between-subjects ANOVA

To Begin SPSS

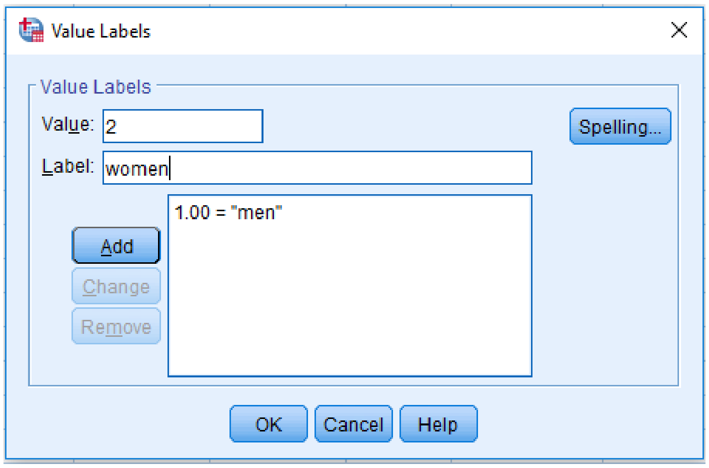
Step 1 Activate the program, close the central window, and click on the **Variable View** option at the bottom left of the window.

Step 2 Click on the first empty cell under the column heading 'Name'. You now type the name of the first variable for which you have data. We are going to utilize the same data and labels as were previously employed in Table 13.3. These hypothetical data dealt with the question of whether there is a relationship between gender, age and the number of traffic tickets received. We have called these variables 'Gender', 'Age' and 'Data'. Therefore, type 'Gender' in the first empty cell under 'Name'.

Step 3 Click on the first empty 'cell' under the column heading 'Label'. In this cell you can type a more extensive description of your variable. In our case, type 'Gender of Subject'.

Step 4 Click on the first empty 'cell' under the column heading 'Values'. A box will appear. In the blank space to the right of 'Value', type the number '1' and then 'men' in the blank space to the right of 'Label'. Finally, click on 'Add'. Your label for a value of 1 will appear in the large white region in the center of the window. Now repeat the above steps in this section for the value '2', which is given the label 'women' (Figure 13.6). Click 'Add' and then click on 'OK'.

Figure 13.6 The Value Labels Window



Value: 2

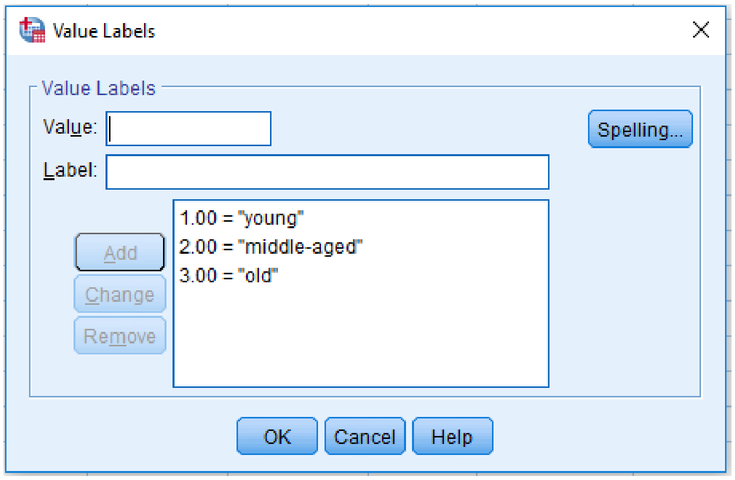
Label: women

1.00 = "men"

Step 5 Click on the first empty ‘cell’ under the column heading ‘Measure’. As we are dealing with labels for groups, select ‘Nominal’ as is shown in the first row of Figure 13.8.

Step 6 Repeat Steps 2 – 4 (for the second IV) except that you type ‘Age’ in the first empty cell under ‘Name’, ‘Age of Subject’ for the label and you now have three values; ‘young’, ‘middle-aged’ and ‘old’ (Figure 13.7).

Figure 13.7 The Second Value Labels Window

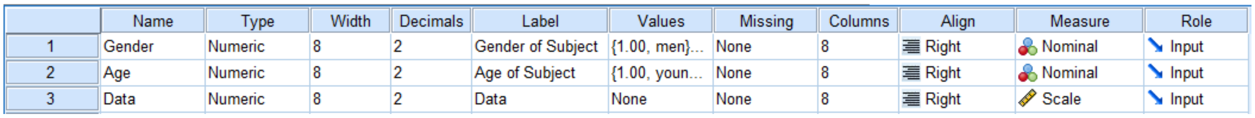


1.00 = "young"
2.00 = "middle-aged"
3.00 = "old"

Step 7 As before, select ‘Nominal’ in the column under the column heading ‘Measure’ as we dealing with labels for groups. The result is shown in the second row of Figure 13.8.

Step 8 We will now repeat the above steps, now for the DV. Type ‘Data’ in the first empty cell under ‘Name’ and for the label. Finally, select ‘Scale’ in the column under the column heading ‘Measure’ as we have ratio data. The result is shown in the third row of Figure 13.8.

Figure 13.8 The Completed Variable View Window



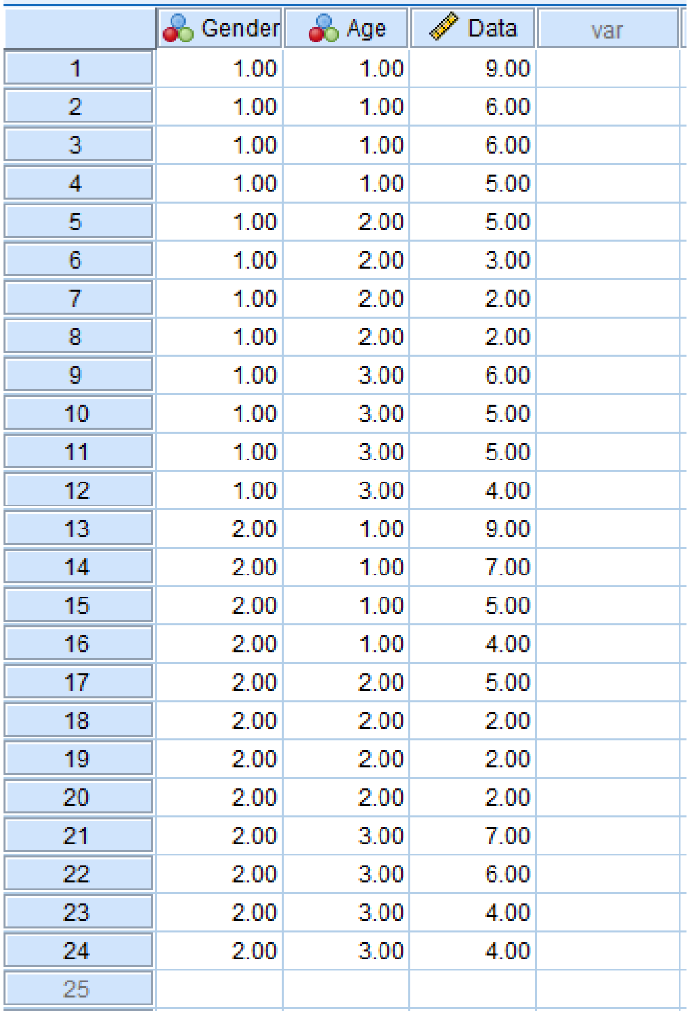
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Gender	Numeric	8	2	Gender of Subject	{1.00, men}...	None	8	Right	Nominal	Input
2	Age	Numeric	8	2	Age of Subject	{1.00, youn...	None	8	Right	Nominal	Input
3	Data	Numeric	8	2	Data	None	None	8	Right	Scale	Input

To Enter Data In SPSS

Step 9 Click on the '**Data View**' option at the lower left corner of the window. The variables 'Gender', 'Age' and 'Data' will be present.

Step 10 For each of the men, type the value '1' in the column 'Gender'. Then in the column 'Age' type '1' if they were young, '2' if they were middle-aged and '3' if they were old. Finally, type the number of tickets each subject received in the third column, 'Data'. Continue by entering '2' for each of the women, along with the value associated with their age and the number of tickets they received (Figure 13.9).

Figure 13.9 The Completed Data Set



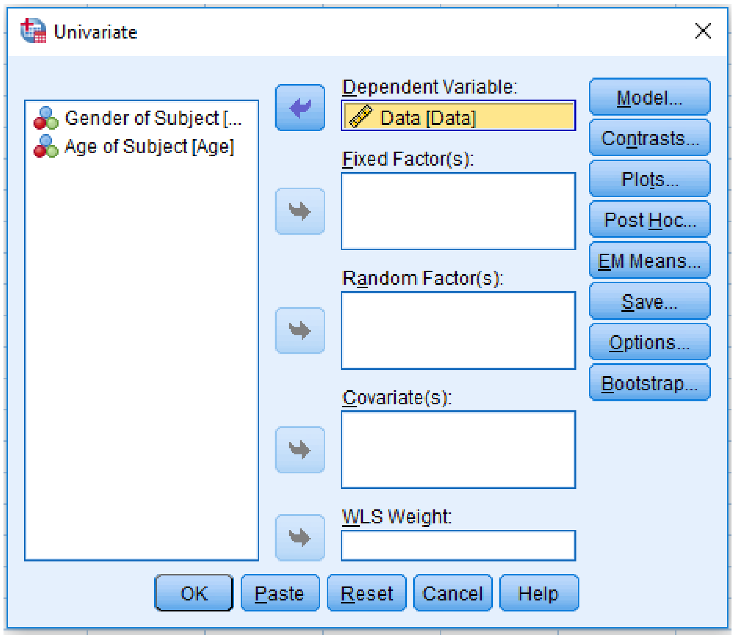
	Gender	Age	Data	var
1	1.00	1.00	9.00	
2	1.00	1.00	6.00	
3	1.00	1.00	6.00	
4	1.00	1.00	5.00	
5	1.00	2.00	5.00	
6	1.00	2.00	3.00	
7	1.00	2.00	2.00	
8	1.00	2.00	2.00	
9	1.00	3.00	6.00	
10	1.00	3.00	5.00	
11	1.00	3.00	5.00	
12	1.00	3.00	4.00	
13	2.00	1.00	9.00	
14	2.00	1.00	7.00	
15	2.00	1.00	5.00	
16	2.00	1.00	4.00	
17	2.00	2.00	5.00	
18	2.00	2.00	2.00	
19	2.00	2.00	2.00	
20	2.00	2.00	2.00	
21	2.00	3.00	7.00	
22	2.00	3.00	6.00	
23	2.00	3.00	4.00	
24	2.00	3.00	4.00	
25				

To Conduct A Two-way Between-subjects ANOVA

Step 11 Click the cursor on '**Analyze**' along the row of SPSS commands above the data you entered, then move to '**General Linear Model**', then click on '**Univariate**'.

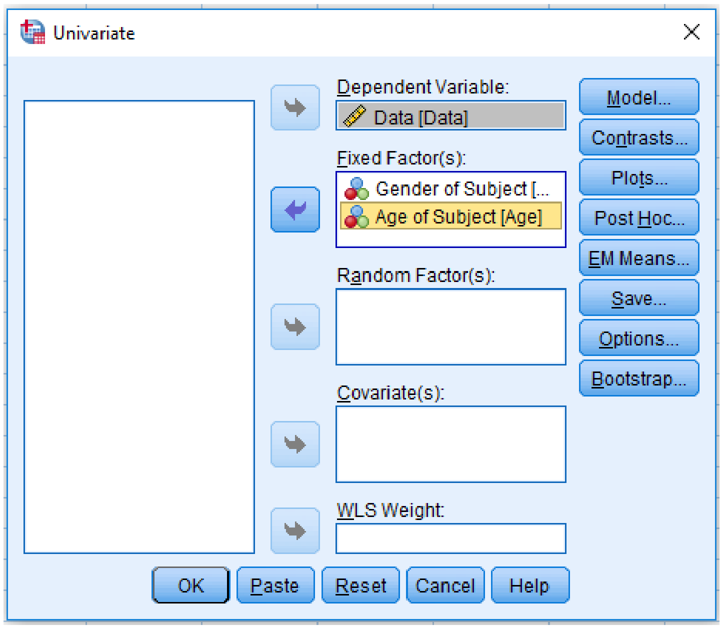
Step 12 A new window will appear. You must now identify the DV and the two IVs (each IV is called a Factor). In our case, 'Data' is the label of the DV. This is indicated by moving 'Data' to the box under 'Dependent Variable' by clicking on 'Data' and then clicking on the top arrow in the box. The result is shown in Figure 13.10. Then click on 'Gender of Subject' and move it to the box under 'Fixed Factor(s)' by clicking on the second arrow. Next, click on 'Age of Subject' and move it to the box under 'Fixed Factor(s)' by clicking on the second arrow. The result will be that each label will move to the appropriate box on the right-hand side of the window, as is shown in Figure 13.11. Then click on '**Post Hoc**' which is located in the column at the far right of the window.

Figure 13.10 Defining Variables



Dependent Variable: Data [Data]

Figure 13.11 Conclusion of Defining Variables



Dependent Variable: Data [Data]

Fixed Factor(s): Gender of Subject [...
Age of Subject [Age]

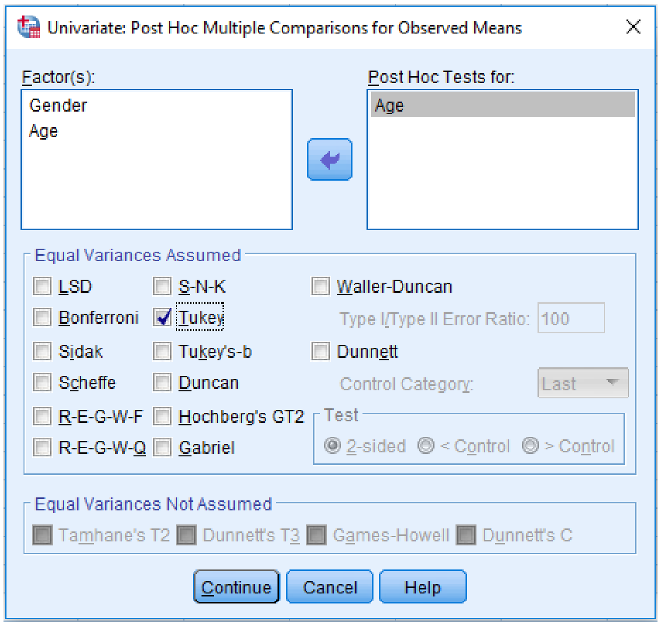
Random Factor(s):

Covariate(s):

WLS Weight:

Step 13 A new window will appear with Gender and Age identified. As the variable 'Gender' has only two levels (men and women) there is no need for a post hoc test. However, the variable 'Age' has three levels (young, middle-aged, and old) so we do conduct a post hoc test in this case. We click on 'Age' and then copy it to the right-hand box by clicking on the arrow. The window will then provide a number of statistical options that are available with SPSS. In this course we will limit ourselves to just the Tukey HSD test. Click on the box next to '**Tukey**' (Figure 13.12). Then click on '**Continue**'.

Figure 13.12 Defining the Post Hoc Test


Univariate: Post Hoc Multiple Comparisons for Observed Means
Continue
Cancel
Help

Factor(s): Gender
Age

Post Hoc Tests for: Age

Equal Variances Assumed

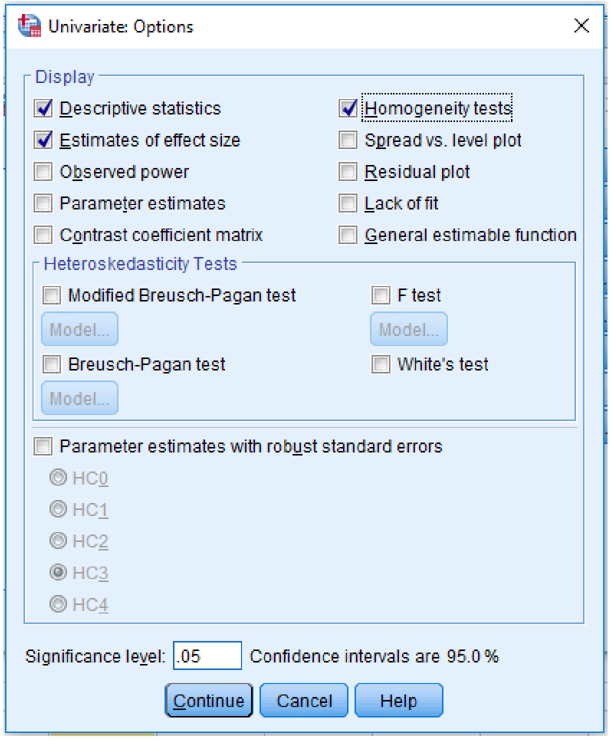
□ LSD □ S-N-K □ Waller-Duncan
□ Bonferroni ☑ Tukey Type I/Type II Error Ratio: 100
□ Sidak □ Tukey's-b □ Dunnett
□ Scheffe □ Duncan Control Category: Last
□ R-E-G-W-F □ Hochberg's GT2 Test
□ R-E-G-W-Q □ Gabriel ● 2-sided ○ < Control ○ > Control

Equal Variances Not Assumed

□ Tamhane's T2 □ Dunnett's T3 □ Games-Howell □ Dunnett's C

Step 14 You have returned to Figure 13.11. In order to obtain descriptive statistics, a measure of effect size, as well as conduct Levene's test for equality (homogeneity) of variances click on '**Options**' which is located in the column at the far right of the window, and then check '**Descriptive statistics**', '**Estimates of effect size**', and '**Homogeneity tests**' as is shown in Figure 13.13. Then click on '**Continue**'.

Figure 13.13 Specifying Descriptive Statistics, Estimates of Effect Size, and Levene's Test



Univariate: Options

Display

☑ Descriptive statistics
☑ Estimates of effect size
☐ Observed power
☐ Parameter estimates
☐ Contrast coefficient matrix
☑ Homogeneity tests
☐ Spread vs. level plot
☐ Residual plot
☐ Lack of fit
☐ General estimable function

Heteroskedasticity Tests

☐ Modified Breusch-Pagan test
☐ F test
☐ Breusch-Pagan test
☐ White's test

☐ Parameter estimates with robust standard errors

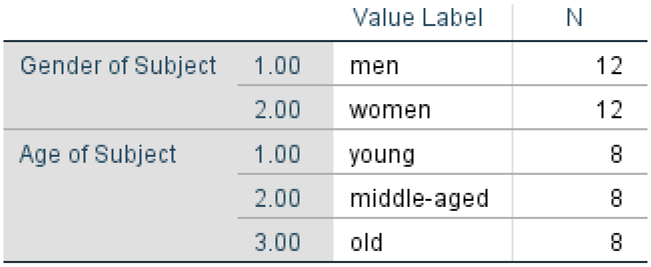
○ HC0
○ HC1
○ HC2
● HC3
○ HC4

Significance level: .05 Confidence intervals are 95.0 %

Step 15 You will have been returned to Figure 13.11. Now click on '**OK**'. SPSS provides descriptive statistics and Levene's test of equality of variances, and calculates the desired two-way ANOVA with partial eta squared and the Tukey HSD post hoc test. Specifically, Table 13.16 provides a count of the number of subjects for each level of the two independent variables 'Gender' and 'Age'. Table 13.17 gives useful descriptive statistics, including means and standard deviations, for each of the levels of our two independent variables as well as for the six cells. It closely parallels, but is more comprehensive, than Table 13.3, which we created by hand. SPSS next provides the output for Levene's test of equality of variances (Table 13.18). There are a number of choices. 'Based on Median' has generally been found to be a good option. As the value of the significance (p-value) is .659, and is thus greater than .05, we maintain the assumption that the samples are drawn from populations with equal variances and we continue to the ANOVA summary table (Table 13.19). We can ignore the first two rows as well as the next to last row. And what we have called 'Within', SPSS labels 'Error'. Otherwise, it is the same outcome as we found earlier, except for rounding error, with our hand calculations (Table 13.5). And the value of the partial eta squared (last column in

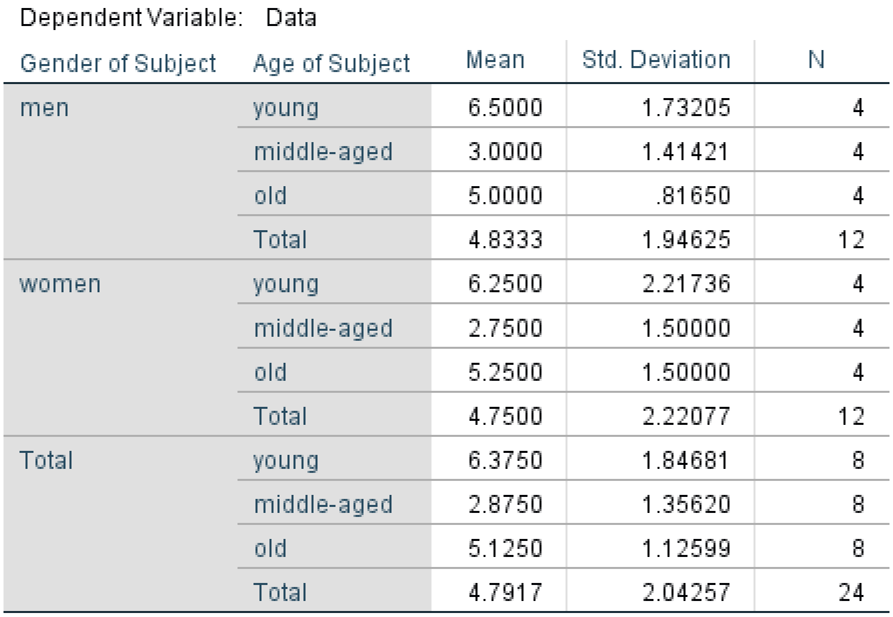
Table 13.19) for Age (Factor B) is the same as we previously obtained except we rounded to two places. (Note that SPSS calculates effect sizes for all of the F ratios, not just those that are statistically significant.) The ANOVA summary table is followed by Tables 13.20 and 13.21 which show the results of the Tukey HSD post hoc test for our significant main effect of Age. The results of the Tukey HSD post hoc test (Table 13.20) correspond to what we calculated previously, though the presentation is different, and 95% confidence intervals are included. (Note that in Table 13.20 the * in the column 'Mean Difference' indicates that the comparison is statistically significant. The final table of the SPSS output (Table 13.21) provides an alternative way of presenting the results of the Tukey HSD post hoc test. It indicates which comparisons differ by the column in which the means are listed. Thus Table 13.21 shows that the middle-age group differed from the young and the old groups, but the young and the old groups did not differ from each other. We can ignore the last row of Table 13.21.

Table 13.16 SPSS Output; Between-Subjects Factors



		Value Label	N
Gender of Subject	1.00	men	12
	2.00	women	12
Age of Subject	1.00	young	8
	2.00	middle-aged	8
	3.00	old	8

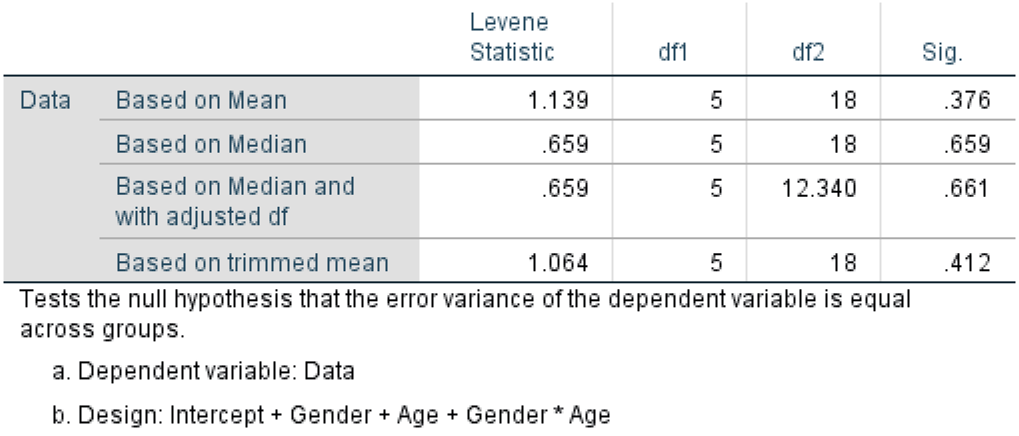
Table 13.17 SPSS Output; Descriptive Statistics



Dependent Variable: Data

Gender of Subject	Age of Subject	Mean	Std. Deviation	N
men	young	6.5000	1.73205	4
	middle-aged	3.0000	1.41421	4
	old	5.0000	.81650	4
	Total	4.8333	1.94625	12
women	young	6.2500	2.21736	4
	middle-aged	2.7500	1.50000	4
	old	5.2500	1.50000	4
	Total	4.7500	2.22077	12
Total	young	6.3750	1.84681	8
	middle-aged	2.8750	1.35620	8
	old	5.1250	1.12599	8
	Total	4.7917	2.04257	24

Table 13.18 SPSS Output; Levene's Test of Equality of Error Variances



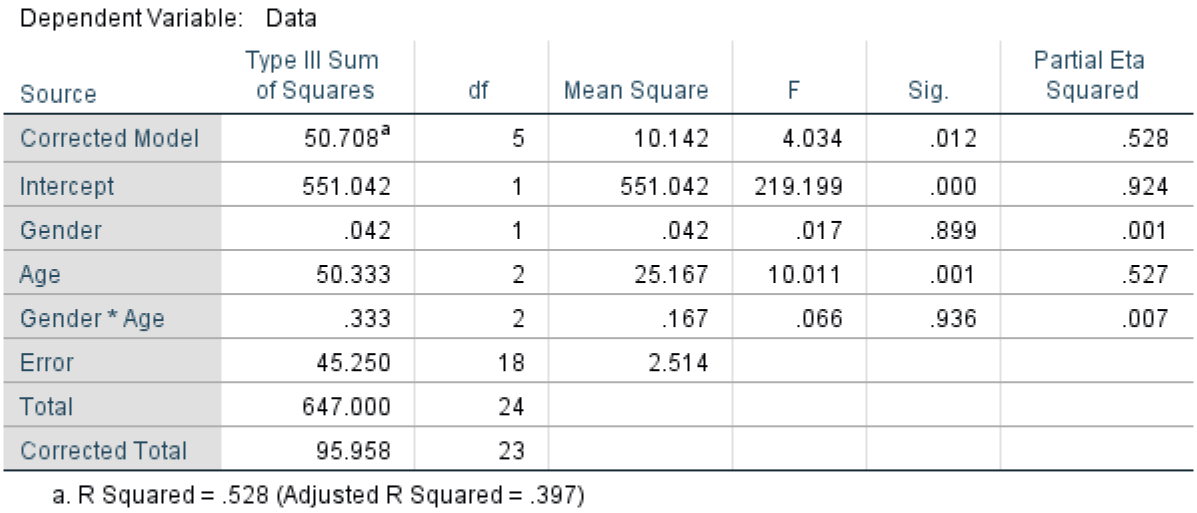
		Levene Statistic	df1	df2	Sig.
Data	Based on Mean	1.139	5	18	.376
	Based on Median	.659	5	18	.659
	Based on Median and with adjusted df	.659	5	12.340	.661
	Based on trimmed mean	1.064	5	18	.412

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.

a. Dependent variable: Data

b. Design: Intercept + Gender + Age + Gender * Age

Table 13.19 SPSS Output; ANOVA Table, Tests of Between-Subjects Effects

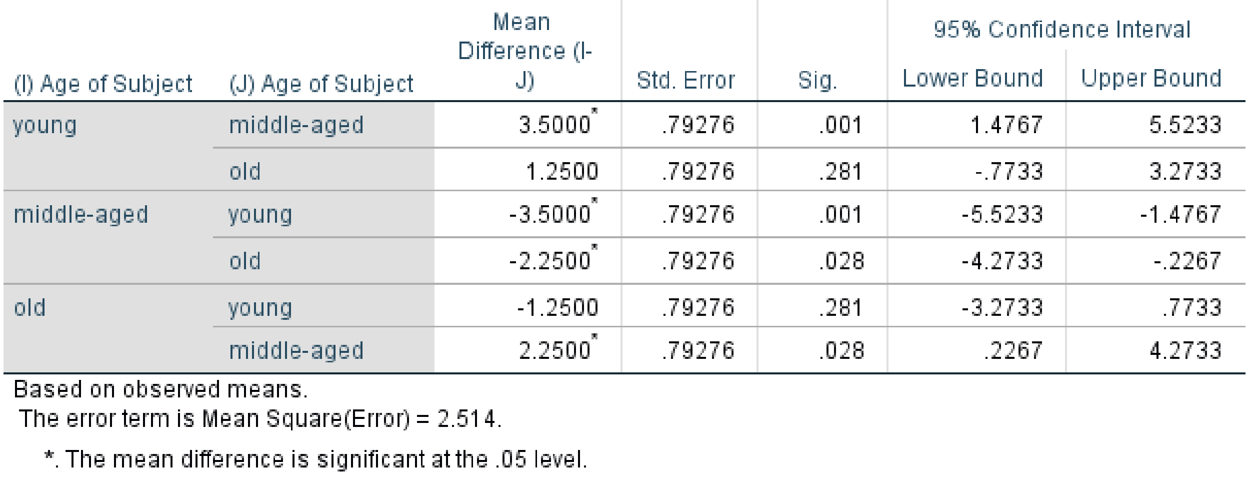


Dependent Variable: Data

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	50.708[a]	5	10.142	4.034	.012	.528
Intercept	551.042	1	551.042	219.199	.000	.924
Gender	.042	1	.042	.017	.899	.001
Age	50.333	2	25.167	10.011	.001	.527
Gender * Age	.333	2	.167	.066	.936	.007
Error	45.250	18	2.514			
Total	647.000	24				
Corrected Total	95.958	23				

a. R Squared = .528 (Adjusted R Squared = .397)

Table 13.20 SPSS Output; Tukey's HSD Post Hoc for Age of Subject

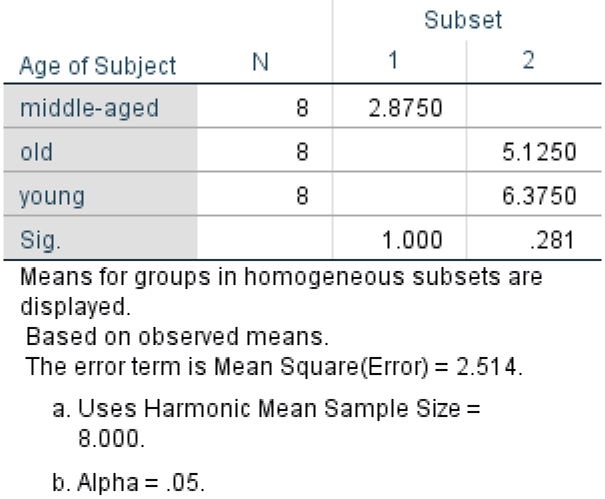


(I) Age of Subject	(J) Age of Subject	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
young	middle-aged	3.5000*	.79276	.001	1.4767	5.5233
	old	1.2500	.79276	.281	-.7733	3.2733
middle-aged	young	-3.5000*	.79276	.001	-5.5233	-1.4767
	old	-2.2500*	.79276	.028	-4.2733	-.2267
old	young	-1.2500	.79276	.281	-3.2733	.7733
	middle-aged	2.2500*	.79276	.028	.2267	4.2733

Based on observed means.
The error term is Mean Square(Error) = 2.514.

*. The mean difference is significant at the .05 level.

Table 13.21 SPSS Output; Alternative presentation of Tukey HSD Post Hoc for Age of Subject



		Subset	
Age of Subject	N	1	2
middle-aged	8	2.8750	
old	8		5.1250
young	8		6.3750
Sig.		1.000	.281

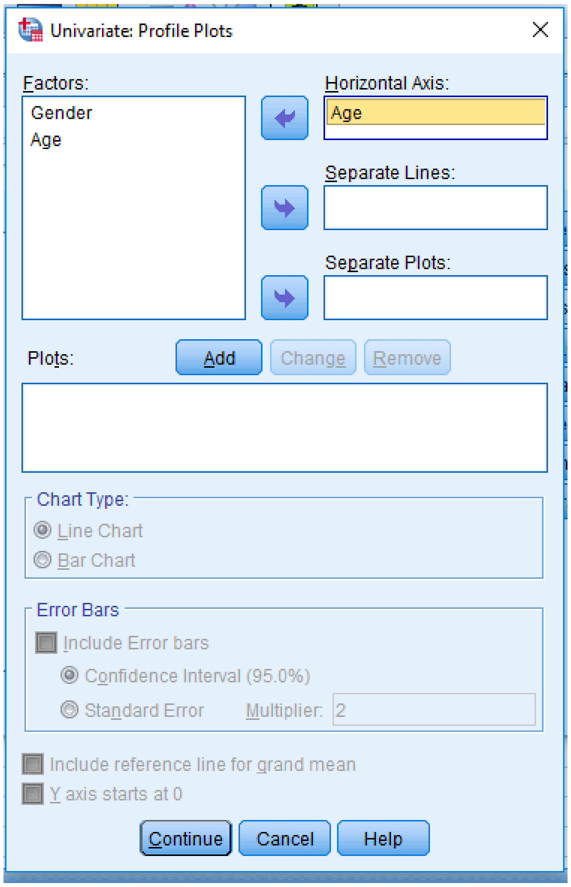
Means for groups in homogeneous subsets are displayed.
Based on observed means.
The error term is Mean Square(Error) = 2.514.

a. Uses Harmonic Mean Sample Size = 8.000.

b. Alpha = .05.

Step 16 As the main effect for Age was found to be significant (Table 13.19) you may want to create a graph to assist in visualizing the results. The means for the young, middle-aged and old groups are available in Table 13.17 and thus a graph, such as is shown in Figure 13.4, can easily be created by hand. Alternatively, you could return to Step 11. You will then see Figure 13.11. Click on '**Plots**' which is located in the column at the far right of the window. Then copy the variable 'Age' across to the right so it is listed under 'Horizontal Axis' (Figure 13.14).

Figure 13.14 Creating a Plot of the Significant Main Effect for Age



Factors:
Gender
Age

Horizontal Axis: Age

Separate Lines:

Separate Plots:

Plots:

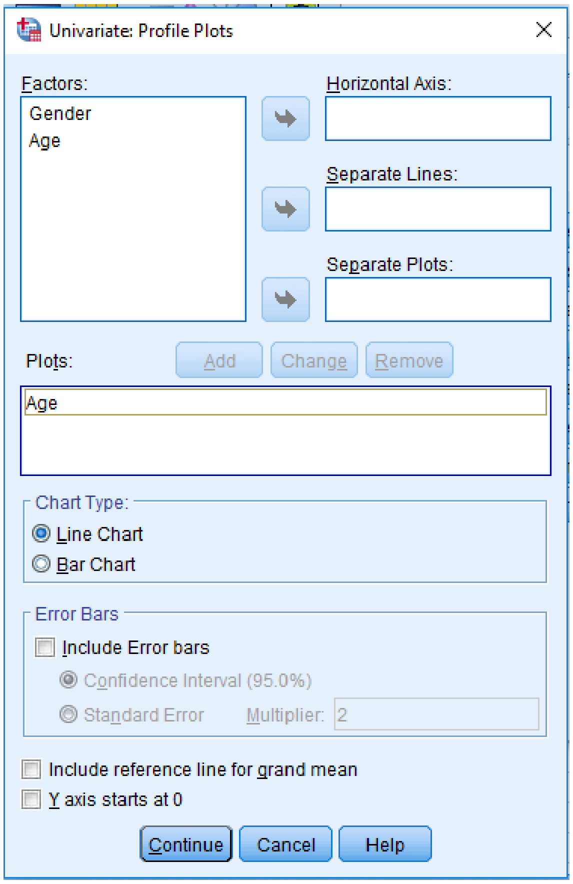
Chart Type:
● Line Chart
○ Bar Chart

Error Bars
□ Include Error bars
● Confidence Interval (95.0%)
○ Standard Error Multiplier: 2

□ Include reference line for grand mean
□ Y axis starts at 0

Step 17 Now click '**Add**'. The result is shown in Figure 13.15.

Figure 13.15 Continuing to Creat a Plot of the Significant Main Effect for Age



Factors:
Gender
Age

Plots:
Age

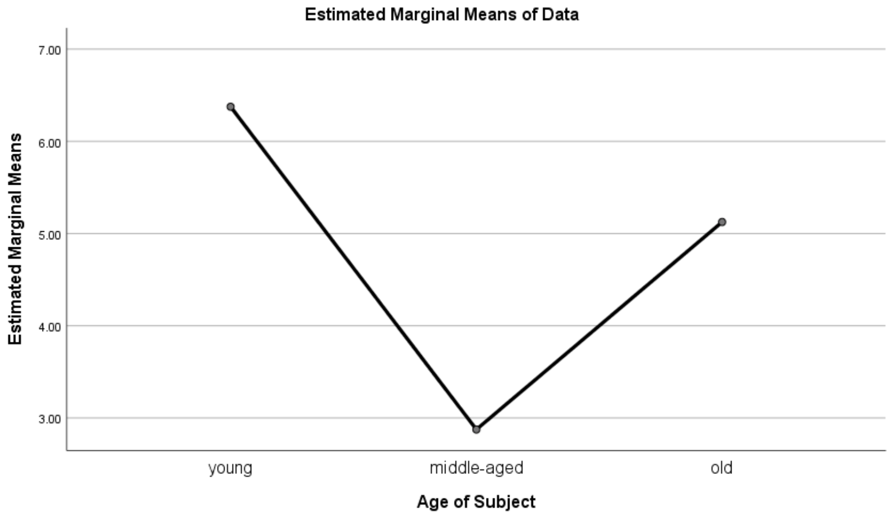
Chart Type:
● Line Chart
○ Bar Chart

Error Bars
□ Include Error bars
● Confidence Interval (95.0%)
○ Standard Error Multiplier: 2

□ Include reference line for grand mean
□ Y axis starts at 0

Step 18 Finally click '**Continue**' and '**OK**'. The data analysis will be given again followed by the plot of our significant main effect for age, which is shown in Figure 13.16. Note that it is similar to the plot we made by hand previously (Figure 13.4) except that the Y-axis does not begin at 0.

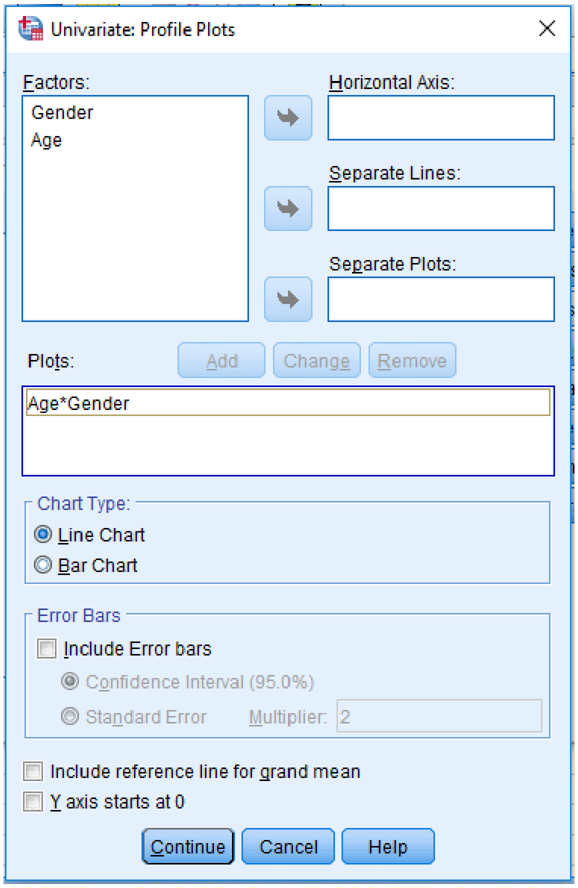
Figure 13.16 Plot of the Significant Main Effect for Age



Step 19 A graph of the nonsignificant interaction could also be created by hand from the means for each cell, which are also available in Table 13.17. Alternatively, you could return to Step

11. You will then see Figure 13.11. Click on ‘**Plots**’ which is located in the column at the far right of the window. Then copy the variable with more levels, in this case ‘Age’ across to the right so it is listed under ‘Horizontal Axis’, and copy the variable with fewer levels, in this case ‘Gender’ across to the right so it is listed under ‘Separate Lines’. Then click on ‘**Add**’. The result is shown in Figure 13.17.

Figure 13.17 Creating a Plot of the Nonsignificant Interaction



Univariate: Profile Plots

Factors:
Gender
Age

Horizontal Axis:

Separate Lines:

Separate Plots:

Plots:
Age*Gender

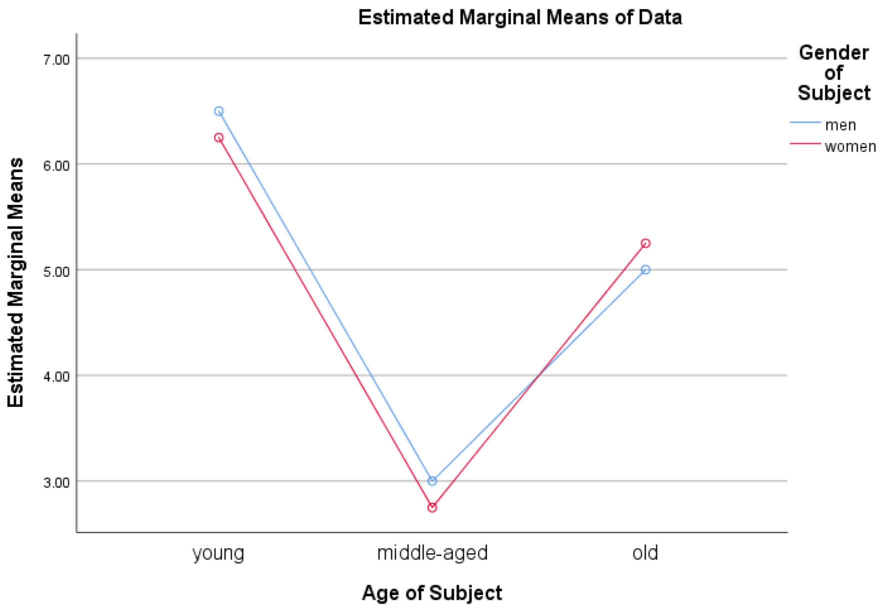
Chart Type:
● Line Chart
○ Bar Chart

Error Bars
□ Include Error bars
● Confidence Interval (95.0%)
○ Standard Error Multiplier: 2

□ Include reference line for grand mean
□ Y axis starts at 0

Step 20 Finally, click ‘**Continue**’ and ‘**OK**’. The data analysis will be given again followed by the plot of our nonsignificant interaction, which is shown in Figure 13.18.

Figure 13.18 Plot of the Interaction, which was not Significant



Step 21 Exit SPSS. There is no need to save the output or data file.

To confirm that you understand how to use SPSS, I suggest you redo the second example of a two-way between-subjects ANOVA that was reviewed in this chapter. The data are presented in Table 13.6. Remember, SPSS does not calculate post hoc comparisons for interactions.

SPSS Problems – Chapter 13

For questions 17 – 20, we are adding a constant, in this case 10, to every score in the noisy condition of the data used for questions 14 – 16. Compare your answers with the answers you found previously.

		Level of Background Noise		
		Quiet	Moderate	Noisy
Gender	Women	9	7	16
		10	9	18
		8	8	20
	Men	13	13	17
		12	11	21
		14	12	22

17. What is the p-value (Sig.) of Levene's test, based on the median, for equality of variances?
    a. .003
    b. .048
    c. .547
    d. .767

18. In the ANOVA table what is the F for gender?
    a. 58.4
    b. 0.8

c. 20
d. 1.0

19. What is the F for background noise?
a. 58.4
b. 0.8
c. 20
d. 1.0

20. What is the F for the interaction of gender X background?
a. 58.4
b. 0.8
c. 20
d. 1.0

For questions 21 – 24 we are subtracting 5 from the scores of the first woman subject in each condition. Compare your answers with the answers you found for questions 17 – 20.

		Level of Background Noise		
		Quiet	Moderate	Noisy
		4	2	11
	Women	10	9	18
		8	8	20
Gender				
		13	13	17
	Men	12	11	21
		14	12	22

21. What is the p-value (Sig.) of Levene's test, based on the median, for equality of variances?
a. .006
b. .043
c. .655
d. .784

22. In the ANOVA table what is the F for gender?
a. 0.218
b. 15.927
c. 12.273
d. 18.472

23. What is the F for background noise?
a. 0.218
b. 15.927
c. 12.273
d. 18.472

24. What is the F for the interaction of gender X background?
a. 0.218
b. 15.927
c. 12.273
d. 18.472

PROCEDURES THAT ARE BOTH DESCRIPTIVE AND INFERENTIAL

Chapter 14
Identifying Associations with Interval or Ratio Data: The Pearson Correlation And Regression

"Statistics is the grammar of science."

Karl Pearson

Introduction

It was pointed out previously that in the broadest sense statistical analysis is undertaken to achieve one of two goals. The goals are to describe your data more clearly, or to make inferences based upon your data. The first chapters of this book dealt with the statistical procedures that are employed when describing data. Together they are called, appropriately, descriptive statistics. Then we introduced the concept of inferential statistics. We noted that inferential statistics are the procedures we use to predict whether a relationship observed in a sample(s) is also likely to exist in a population(s). And we discussed that inferential statistical procedures address two broad questions, is there a difference or is there an association between the variables? (This distinction is not as clear with nominal data. Thus in Chapter 8 we noted that the chi-square test of independence can be used to address either question.)

We have just completed our review of some of the most common statistical procedures used for examining whether a *difference* observed in the data is likely to generalize to the corresponding population(s). In the current chapter we review a commonly employed procedure that is utilized to identify whether an *association* exists among two variables. This procedure can be used to describe a relationship in a sample, in which case it is being used as a descriptive statistic. Alternatively, and more commonly, we can also use this same procedure to test whether a relationship that is observed in a sample is likely to also exist in the population from which the sample was drawn. In this case we are dealing with inferential statistics. Same statistical procedure, different goal.

When we find there is an association we are indicating that two variables are not independent. In other words, they are related or **covary**. And with a **correlation** we indicate the extent to which they are associated. For instance, from casual observation it appears that a person's weight is related to how tall they are. With correlation we can specify the extent of this relationship.

<u>Covary</u> – If knowledge of how one variable changes assists you in predicting the value of another variable, the two variables are said to covary.

<u>Correlation</u> – A measure of the degree of association among variables. A correlation

indicates whether a variable changes in a predicable manner as another variable changes.

Table 14.1, which is a part of the Overview Table (Appendix L), indicates that the specific correlational procedure that is used will depend upon the type of data that is being collected. Specifically, if both variables consist of nominal data you use the Phi correlation, which can be written as Phi r or simply Phi (reviewed in Appendix B). If both variables consist of ordinal data you use the Spearman correlation, which is commonly called the Spearman r (reviewed in Appendix C). And if you are dealing with two interval or ratio variables you would employ the Pearson correlation, which is commonly called the Pearson r. (In statistics, the letter 'r' indicates a correlation.) The Pearson r is the focus of the initial portion of this chapter. (It is underlined in Table 14.1.) Calculation of Phi r, Spearman r or Pearson r results in a **correlation coefficient**, a single number that indicates the degree to which two variables are related.

Correlation coefficient – A single number that indicates the degree to which two variables are related.

Table 14.1 Statistical Procedures used with Association Designs

	Type of Data		
	Nominal (Frequency)	Ordinal (Ranked)	Interval/Ratio (Continuous Measure)
Research Question			
Association:	Chi-Square Test of Independence		
Correlation:	*Phi r*[a]	*Spearman r*[b]	Pearson r *Multiple Correlation*[c]
Regression:			Regression *Multiple Regression*[c]

Italicized items are reviewed in the following appendixes:

a. Appendix B
b. Appendix C
c. Appendix D

Pearson Correlation

The Pearson correlation, or Pearson r, is also sometimes called the Pearson product–moment correlation coefficient. It is undoubtedly the most commonly used form of correlation.

With the Pearson r we use the symbol ρ_{XY} or simply ρ (**rho**, pronounced row) to indicate the population correlation between two variables X and Y, and r_{XY} or simply r to indicate a correlation between X and Y found in a sample.

__Rho__ (ρ) – Symbol used for the population correlation.

In the case of the Pearson r, the two variables of interest are measured at the interval or ratio level. As the Pearson r is a measure of linear relationship it only provides an accurate indication of the magnitude of an association if the two variables have a straight-line relationship between them. If there is not a linear relationship between the variables the Pearson r will underestimate the true degree of the association. In this case the data could be transformed (refer to a more advanced text). Alternatively, the data could be converted to ranks and the Spearman r correlation would then be calculated (This procedure is reviewed in Appendix C).

With the Pearson r, the sign of the correlation (positive or negative) indicates the direction of the relation. A **positive correlation** indicates that as one variable increases, so does the other (Figure 14.1). For instance, in general those students who study more get higher grades. A positive correlation also indicates that as one variable decreases, so does the other. In other words, those students who study less tend to get lower grades. With a **negative correlation**, as one variable increases, the other decreases (Figure 14.2). An example of a negative correlation would be the total mileage of a used car and how much it is worth. In general, the more miles the car has been driven, the less it is worth. That is why some unscrupulous individuals used to roll back odometers before selling their cars. The cars had gone just as many miles, but the buyers were not aware of this and paid more than they would have if they had known the true situation.

The magnitude of the Pearson correlation, ignoring the sign, indicates the size of the relationship. For instance, the largest absolute value for a Pearson r is 1, which corresponds to a value of +/–1. If the Pearson r is equal to +/–1 there is a perfect association among the variables. In other words, if you know the value of one variable you can predict the value of the other variable perfectly, without any error. And a graph or scatter plot would show that all of the data points fall along a straight line. If the correlation is +1, the line would rise to the right (Figure 14.1). If the correlation is –1, the line would rise to the left (Figure 14.2). In the real world, correlations of +/–1 are unlikely to occur. Instead we find more modest correlations with values such as +.32 or –.57. When the magnitude is between +1 and -1 the data points would not all fall directly on a straight line (Figure 14.3), and while knowing the value of one variable will be of some assistance in predicting the value of the other variable, the predictions will not be perfect. An example of this is the weather forecast on the nightly news. The forecast is not always correct, but we pay attention because it is much more accurate than simply guessing. If the Pearson r is equal to 0, the variables are unrelated and knowing the value of one variable does not assist in predicting the value of the other variable (Figure 14.4).

<u>Positive correlation</u> – A relationship between two variables in which as one variable increases in value, so does the other variable. Also, as one variable decreases in value, so does the other.

<u>Negative correlation</u> – A relationship between two variables in which as one variable increases in value, the other variable decreases in value. Also, as one variable decreases in value, the other increases in value.

Figure 14.1 A Positive Correlation. More Specifically, a Correlation of +1

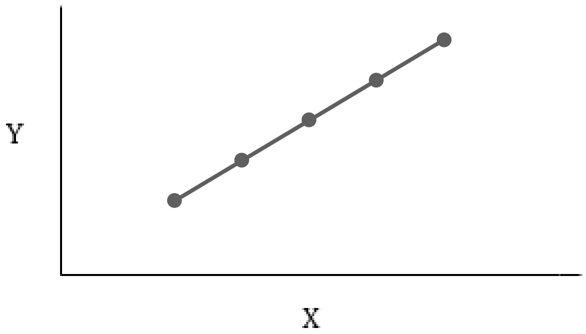


Figure 14.2 A Negative Correlation. More Specifically a Correlation of –1

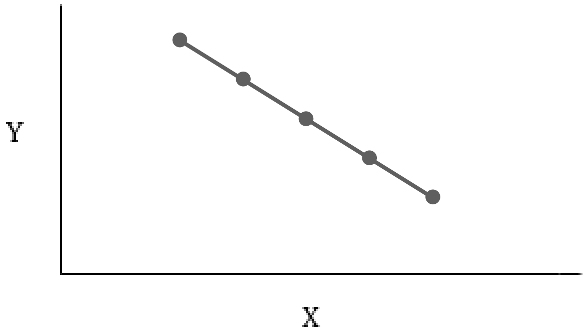


Figure 14.3 An Intermediate, Positive Correlation

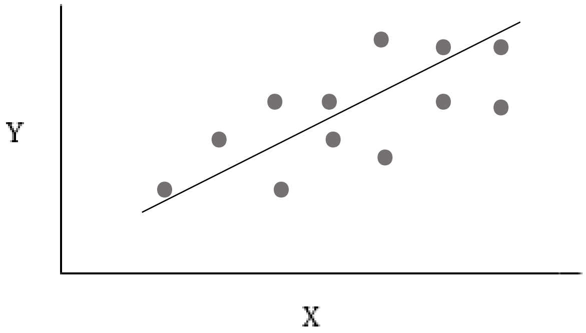
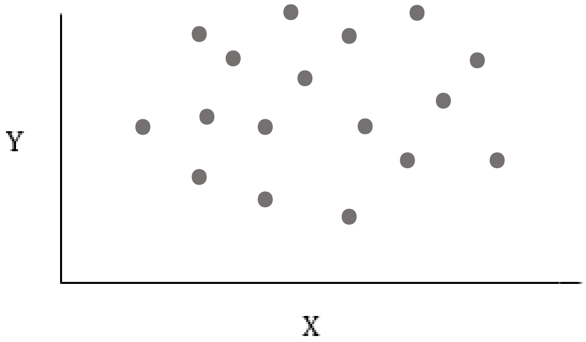


Figure 14.4 A Correlation of 0



Thus with a Pearson correlation the sign indicates the direction of the association while the magnitude indicates the degree to which the two variables are related, in other words how well we can predict from one variable to another. What a Pearson correlation does not provide is the actual equation that would permit a researcher to predict the value of one variable when the value of the other variable is known. In other words, with just a statistically significant Pearson correlation researchers know that a prediction can be made and how well it can be made, but they do not know what the actual prediction would be. In order to make a prediction we employ a closely related statistical procedure called **regression**, which will be reviewed later in this chapter. (This procedure is underlined in Table 14.1.)

Regression – Procedure researchers use to develop an equation that permits the prediction of the value of one variable of a correlation if the value of the other variable is known.

We have often discussed the concepts of Type I and Type II errors in this book. Each of these errors can also be made with a correlational study. For instance, if we conclude, based upon our samples, that a correlation exists in a population when in fact there is no such correlation, we have made a Type I error. In other words, we have rejected the null hypothesis which is usually that the population correlation (ρ) is 0 even though the null hypothesis is actually true. On the other hand, if we conclude that there is no correlation among the variables in the population, when in fact there is, we have made a Type II error. In this case we have failed to reject the null hypothesis even though it is false.

In a correlational study no independent variable is manipulated by the researcher and there is no control group. Independent variables and control groups, as you have learned, are characteristics of experiments. They do not occur with correlational studies. Instead, in a correlational study the researcher records information concerning naturally occurring variables and later determines whether these variables are associated. Correlational studies are generally easier to conduct than experiments and numerous variables can be examined quickly. The studies are, in this sense, efficient. However, an important limitation of correlational studies is that their results do not justify coming to a strong, cause-and-effect conclusion. For instance, the initial scientific findings linking smoking with cancer were based solely on correlational studies. It quickly was established that an association or linkage existed between smoking and experiencing certain types of cancer. The government responded with warning labels on packages of cigarettes. These labels, however, were much weaker than the current ones since the original, correlational studies did not warrant the current stronger cause-and-effect wording.

Conducting A Pearson Correlation

Our first example of a Pearson r consists of hypothetical quiz and exam scores for seven students taking a course in statistics (Table 14.2). We will set α equal to .05. The null hypothesis (H_0) is that there is no correlation between quiz and exam scores for the population of all students taking a course in statistics. In other words, the null hypothesis is that $\rho_{XY} = 0$ for quiz and exam scores. The alternative hypothesis (H_1) is that $\rho_{XY} \neq 0$. Since no direction is specified for the outcome this is a two-tailed test.

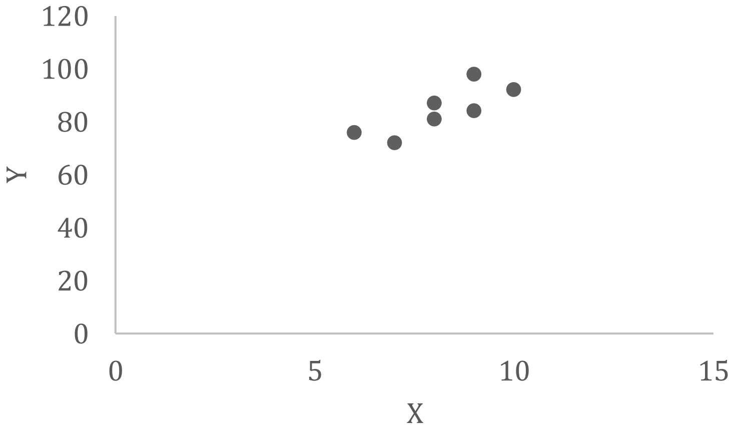
Table 14.2 Example 1: Hypothetical Quiz and Exam Scores

Student	Quiz Score (X)	Exam Score (Y)
1	10	92
2	9	98
3	9	84
4	8	87
5	8	81
6	7	72

7	6	76

A graph of these data suggests that there is a trend such that higher quiz scores are associated with higher exam grades. In other words, the quiz and exam scores appear to vary together (Figure 14.5).

Figure 14.5 Example 1: A Graph of Hypothetical Quiz and Exam Scores



In statistics, the extent to which two variables covary is known as their **covariance**. The equation for the covariance is:

$$\mathrm{cov}_{xy} = \frac{\sum(X - M_X)(Y - M_Y)}{n - 1}$$

where n is the number of *pairs* of scores.

The equation indicates that each value for X is converted into a deviation from its mean. Similarly, each corresponding value for Y is converted into a deviation from its mean. Then each pair of deviations is multiplied together. Next, all of these multiplied deviations are added and, finally, this sum is divided by the number of pairs of scores minus 1.

Covariance – A statistical measure indicating the extent to which two variables vary together.

Upon a closer examination of the equation for the covariance it should be evident that if a value for X is greater than its mean, then $(X – M_X)$ will be positive. Also, if the value for Y that is paired with this X is greater than its mean, then this $(Y – M_Y)$ will also be positive, and the product of these two deviations will thus be positive. Similarly, if a value for X is less than its mean, then $(X – M_X)$ will be negative. Also, if the value for Y that is paired with this X is less than its mean, then this $(Y – M_Y)$ will also be negative, and their product will thus be positive. However, if one of the deviations is positive and the other is negative, then the product will be negative. The sum of all of the products will determine the sign of the covariance, and thus will indicate the direction in which the two variables covary.

With our example, the individuals with the highest quiz grades tend to also have the highest exam grades. Thus, the greatest positive deviations for the quizzes will tend to be matched with the greatest positive deviations for the exam scores. Similarly, the individuals with the lowest quiz grades tend to also have the lowest exam grades. Thus, the largest negative deviations for the quizzes will tend to be matched with the largest negative deviations for the exam scores. As was just noted this pairing of positive deviations in one variable with positive deviations in the other, and negative deviations in one variable with negative deviations in the other, will lead to a positive value for the covariance. If the positive deviations for the quizzes had tended to be matched with the negative deviations for the exam grades, the covariance would have a negative value. And you will see shortly that the sign of the covariance determines the sign of the Pearson correlation.

Furthermore, the magnitude of the covariance will be a maximum when the most extreme X and Y scores are paired together. And it will be zero if the two variables are not related and thus do not covary.

Covariation is essential to understanding the Pearson correlation. In order to calculate the value of a Pearson r, all that is further needed is to take the magnitude of the standard deviations for the X and Y variables into account. More specifically, the equation for the Pearson correlation is as follows:

$$r_{XY} = r = \frac{cov_{XY}}{s_X s_Y}$$

Conceptually, this equation indicates that the Pearson r is the ratio of a measure of the degree to which two variables covary (vary together) and a measure of the product of their variabilities.

To use this equation it is necessary to determine the value of the covariance and also to calculate the standard deviation of the X scores and the standard deviation of the Y scores. (There are computational equations that are easier to use with large data sets, but the logic for the specific calculations is then not evident. These equations are provided in Appendix G. However, anyone anticipating calculating a Pearson r for a substantial data set is advised to use a computer package instead.)

Calculating The Covariance

To calculate the covariance we proceed as shown in Table 14.3.

Table 14.3 Example 1: Initial Steps in the Calculation of the Covariance for Quiz and Exam Scores

Student	Quiz (X)	Exam (Y)	$(X – M_X)$	$(Y – M_Y)$	$(X – M_X)(Y – M_Y)$

1	10	92	1.86	7.71	14.34
2	9	98	0.86	13.71	11.79
3	9	84	0.86	–0.29	–0.25
4	8	87	–0.14	2.71	–0.38
5	8	81	–0.14	–3.29	0.46
6	7	72	–1.14	–12.29	14.01
7	6	76	–2.14	–8.29	17.74
	$\sum X = 57$	$\sum Y = 590$	$\sum(X - M_X) \approx 0$	$\sum(Y - M_Y) \approx 0$	$\sum(X - M_X)(Y - M_Y) = 57.71$
	$M_X = 8.14$	$M_Y = 84.29$			

Substituting these values into the equation for the covariance, we have:

$$\text{cov}_{xy} = \frac{\sum(X - M_X)(Y - M_Y)}{n - 1}$$

where n is the number of *pairs* of scores.

$$\text{cov}_{XY} = \frac{57.71}{7 - 1}$$

$$= \frac{57.71}{6}$$

$$= 9.62$$

Determining The Standard Deviations

To find the value of the Pearson correlation we now need to calculate the standard deviations for the X and Y scores (Tables 14.4 and 14.5). Recall that the equation for the standard deviation of a sample is:

$$s_X = \sqrt{\frac{\sum(X - M_X)^2}{n - 1}}$$

And note that the needed mean and deviations come from calculations in Table 14.3.

Table 14.4 Example 1: Calculation of the Standard Deviation of the Quiz Scores (X)

Student	Quiz (X)	$(X - M_X)$	$(X - M_X)^2$
1	10	1.86	3.46
2	9	0.86	0.74
3	9	0.86	0.74
4	8	–0.14	0.02
5	8	–0.14	0.02
6	7	–1.14	1.30
7	6	–2.14	4.58
	$\sum X = 57$	$\sum(X - M_X) \approx 0$	$\sum(X - M_X)^2 = 10.86$
	$M_X = 8.14$		

Substituting the values of $\sum(X - M_X)^2$ and n into the equation for the standard deviation of quiz scores, we have:

$$s_X = \sqrt{\frac{\sum(X - M_X)^2}{n - 1}}$$

$$= \sqrt{\frac{10.86}{7 - 1}}$$

$$= \sqrt{\frac{10.86}{6}}$$

$$= \sqrt{1.81}$$

$$= 1.35$$

We now turn to calculating the standard deviation of the Y scores (Table 14.5). Note, however, that the mean of Y and the deviations have already been found (Table 14.3).

Table 14.5 Example 1: Calculation of the Standard Deviation of the Exam Scores (Y)

Student	Exam (Y)	$(Y - M_Y)$	$(Y - M_Y)^2$
1	92	7.71	59.44
2	98	13.71	187.96
3	84	−0.29	0.08
4	87	2.71	7.34
5	81	−3.29	10.82
6	72	−12.29	151.04
7	76	−8.29	68.72
	$\sum Y = 590$	$\sum(Y - M_Y) \approx 0$	$\sum(Y - M_Y)^2 = 485.40$
	$M_Y = 84.29$		

Substituting the values from Table 14.5 into the equation for the standard deviation, we have:

$$s_Y = \sqrt{\frac{\sum(Y - M_Y)^2}{n - 1}}$$

$$= \sqrt{\frac{485.40}{7 - 1}}$$

$$= \sqrt{\frac{485.40}{6}}$$

$$= \sqrt{80.90}$$

$$= 8.99$$

Calculating The Pearson r

We are now in a position to calculate the value of the Pearson correlation. Recall that the equation for the Pearson correlation is:

$$r_{XY} = r = \frac{cov_{XY}}{s_X s_Y}$$

Substituting the values just obtained we have:

$$r_{XY} = r = \frac{9.62}{(1.35)(8.99)}$$

$$= \frac{9.62}{12.14}$$

$$= +.79$$

Interpreting Our Pearson r

The df for the Pearson r = n – 2, where n is the number of *pairs* of scores. In our case, there would be 7 – 2 or 5 df.

Referring to the Pearson r table (Appendix K, Table 6), we find that the critical value for 5 df with α equal to .05 is .75 for a two-tailed test. With a two-tailed test you reject the null and accept the alternative hypothesis if the absolute value of the calculated r is greater than the critical value from the table. As our obtained Pearson r was .79, which is greater than the critical value of .75, we reject the null hypothesis that the population correlation (ρ_{XY}) is equal to 0 and accept the alternative hypothesis that ρ_{XY} is not equal to 0.

Recall that the larger the Pearson correlation, the better we can predict. Our value of the Pearson r is .79. This is a very strong correlation. In other words, if we know a person's quiz grade, we would be able to predict quite well how that person will do on the subsequent exam.

Determining The Effect Size

The strength of the association is determined by finding the square of the correlation, r^2. The square of a correlation is called the **coefficient of determination**. It measures the proportion of variance in one variable that is explained or accounted for by the other variable. In our example, the correlation was equal to .79. Thus r^2 is equal to $.79^2$, which is .62 or 62%. This indicates that knowing a person's quiz score will account for 62% of the variability in predicting their exam score.

<u>Coefficient of determination</u> – The square of the correlation. It indicates the proportion of variability in one variable that is explained or accounted for by the variability in the other variable.

Put another way, there is only 38% of the variability in the exam scores that is *not* accounted for by knowing the quiz scores. This is determined by subtracting 62%, the percentage of the variability that is known, from 100%, which is the total variability. Alternatively, we could express this in terms of a proportion by subtracting .62 from 1.00 to obtain .38. This value, which is the proportion of the variability of one variable *not* explained or accounted for by the variability of the other variable, is called the **coefficient of nondetermination**. For the Pearson r, it is equal to 1 – r^2.

<u>Coefficient of nondetermination</u> – The proportion of the variability of one variable not explained or accounted for by the variability of the other variable. For the Pearson r, it is equal to 1 – r^2.

Reporting The Results Of A Pearson Correlation

To summarize our findings we would indicate the number of degrees of freedom, the calculated value of the Pearson r, and the p-value. Based upon our calculations we would report, "A positive relationship was found between quiz and exam scores ($r(5) = .79$, $p < .05$)." After reporting the significant r, we would then indicate the effect size by saying "r^2 was equal to .62." This example is also completed using SPSS at the end of this chapter. In a paper we would report these more precise findings as well as the p-value obtained when using SPSS, ($r(5) = .80$, $p = .033$). (Note that our p-value of .033 is less than our α of .05 which confirms that we would reject the null hypothesis.) We would then say "r^2 was equal to .64."

A Second Example

It is important to understand that the magnitude of the Pearson r of a small set of data can be dramatically affected by the removal of just one or two extreme scores. This is called **restriction of the range**. For instance, if we omit subjects 6 and 7 from the above data set we would have the data shown in Table 14.6.

<u>Restriction of the range</u> – Reducing the range of values for a variable will reduce the size of the correlation.

Table 14.6 Example 2: Restricted Range of Hypothetical Quiz and Exam Scores

Student	Quiz (X)	Exam (Y)
1	10	92
2	9	98
3	9	84
4	8	87

5	8	81

Calculating The Covariance

The steps involved in calculating the covariance are illustrated in Table 14.7.

Table 14.7 Example 2: Initial Steps in the Calculation of the Covariance for the Restricted Range of Quiz and Exam Scores

Student	Quiz (X)	Exam (Y)	$(X - M_X)$	$(Y - M_Y)$	$(X - M_X)(Y - M_Y)$
1	10	92	1.20	3.60	4.32
2	9	98	0.20	9.60	1.92
3	9	84	0.20	−4.40	−0.88
4	8	87	−0.80	−1.40	1.12
5	8	81	−0.80	−7.40	5.92
	$\sum X = 44$	$\sum Y = 442$	$\sum(X - M_X) = 0$	$\sum(Y - M_Y) = 0$	$\sum(X - M_X)(Y - M_Y) = 12.40$
	$M_X = 8.80$	$M_Y = 88.40$			

Substituting these values into the equation for the covariance, we have:

$$\text{cov}_{xy} = \frac{\sum(X - M_X)(Y - M_Y)}{n - 1}$$

where n is the number of *pairs* of scores.

$$\text{cov}_{XY} = \frac{12.40}{5 - 1}$$

$$= \frac{12.40}{4}$$

$$= 3.10$$

Determining The Standard Deviations

We now need to calculate the standard deviations of the X (Table 14.8) and Y (Table 14.9) scores. Note, however, that the means and deviations have already been calculated (Table 14.7).

Table 14.8 Example 2: Calculation of the Standard Deviation of the Quiz Scores (X) for the Restricted Range of Quiz and Exam Scores

Student	Quiz (X)	$(X - M_X)$	$(X - M_X)^2$
1	10	1.20	1.44
2	9	0.20	0.04
3	9	0.20	0.04
4	8	−0.80	0.64
5	8	−0.80	0.64

$\sum X = 44$ $\quad\quad$ $\sum(X - M_X) = 0$ $\quad\quad$ $\sum(X - M_X)^2 = 2.80$

$M_X = 8.80$

Substituting the values of $\sum(X - M_X)^2$ and n into the equation for the standard deviation of quiz scores, we have:

$$s_X = \sqrt{\frac{\sum(X - M_X)^2}{n - 1}}$$

$$= \sqrt{\frac{2.80}{5 - 1}}$$

$$= \sqrt{\frac{2.80}{4}}$$

$$= \sqrt{0.70}$$

$$= 0.84$$

Table 14.9 Example 2: Calculation of the Standard Deviation of the Exam Scores (Y) for the Restricted Range of Quiz and Exam Scores

Student	Exam (Y)	$(Y - M_Y)$	$(Y - M_Y)^2$
1	92	3.60	12.96
2	98	9.60	92.16
3	84	–4.40	19.36
4	87	–1.40	1.96
5	81	–7.40	54.76
	$\sum Y = 442$ $M_Y = 88.40$	$\sum(Y - M_Y) = 0$	$\sum(Y - M_Y)^2 = 181.20$

Substituting these values into the equation for the standard deviation, we have:

$$s_Y = \sqrt{\frac{\sum(Y - M_Y)^2}{n - 1}}$$

$$= \sqrt{\frac{181.20}{5 - 1}}$$

$$= \sqrt{\frac{181.20}{4}}$$

$$= \sqrt{45.30}$$

$$= 6.73$$

Calculating The Pearson r

We are now in a position to calculate the value of the Pearson correlation. The equation for the Pearson correlation is:

$$r_{XY} = r = \frac{cov_{XY}}{s_X s_Y}$$

Substituting the values just obtained we have:

$$r_{XY} = r = \frac{3.10}{(0.84)(6.73)}$$

$$= \frac{3.10}{5.65}$$

$$= .55$$

Interpreting Our Pearson r

The df for the Pearson r = n – 2, where n is the number of *pairs* of scores. In our case, there would be 5 – 2 or 3 df. The critical value from the Pearson r table (Appendix K, Table 6) with α = .05 is .88 for a two-tailed test. Note that a consequence of losing degrees of freedom is an increase in the critical value from .75 to .88, which will make it more difficult to reject the null hypothesis.

In addition, restricting the range by eliminating the two lowest quiz scores resulted in a drop in the size of the correlation from .79 to .55. This is a substantial decline and the resulting correlation would no longer be statistically significant even without the loss of degrees of freedom. If the outcome were still significant, r_{XY}^2 would have dropped from 62% to 30%, which indicates that the ability to predict exam grades would have been reduced substantially.

Reporting The Results Of A Pearson Correlation

Based upon our calculations using the restricted range of data we would report, "No significant relationship was found between quiz and exam scores ($r(3) = .55$, $p > .05$)." Of course, we should also recognize that the data set was much too small. However, if we were going to put our findings in a publication we would want to utilize a statistical package for our data analysis in order to gain greater precision and to provide a precise p-value. With SPSS we would report ($r(3) = .55$, $p = .336$). Note that the p-value of .336 is greater than the α level of .05.

Purpose And Limitations Of Using The Pearson Correlation

1. *Provides a measure of the association of two interval or ratio variables.* The Pearson correlation provides a measure of the strength and direction of an association between two interval or ratio variables.
2. *Not a measure of cause and effect.* The Pearson r is a type of correlation. Due to a lack of control in a correlational design a researcher is not justified in coming to a cause and effect conclusion.

3. *Be aware of restriction of the range.* A broad range of X and Y values will generally increase the size of the Pearson r. A reduction in the range of these variables will tend to reduce the size of the Pearson r.
4. *Prediction is limited to the range of the original values.* The Pearson r indicates the correlation for a range of X and Y values. The nature of the correlation for these variables is unknown beyond the range included in the original calculations. For instance, if the original correlation of height and weight is based upon people with heights between 5 and 6 feet, then we cannot predict the weight for a person 7 feet tall.

Assumptions Of The Pearson Correlation

1. *Interval or ratio data.* The two variables consist of interval or ratio data.
2. *Data are paired.* The data come as pairs, usually two measures on the same individual.
3. *Linear relationship.* The Pearson correlation assumes that the two variables are linearly related.
4. *Bivariate Normally Distributed.* This assumption is tested by checking that the X and Y variables are each normally distributed. However, so long as there are not outliers the Pearson correlation will be accurate (is robust) so long as the sample size is at least 30.

Progress Check

1. Another term for the square of a correlation is the _____. It indicates the _____ in one variable that is explained or accounted for by variability in the other variable.
2. With ordinal data we would use the _____ as our correlation. If we had two variables measured at the interval or ratio level we would use the _____ correlation.
3. Assume it was reported that the correlation was 0.50 between age and reading ability for students selected from grades 1 to 12. We then measure this relationship for grades 6 through 12 and find a smaller correlation. This is an example of _____.

Answers: 1. coefficient of determination; proportion of variability 2. Spearman r; Pearson 3. restriction of the range

Conclusion Of Correlation

This chapter has introduced the most commonly used correlational technique, the Pearson r. The Pearson r is employed when we have two variables, each consisting of interval or ratio data. However, in the real world numerous variables may be associated together. For instance, many variables are linked with a student's SAT score, including their high school grade average, quality of their teachers, physical characteristics of the high school, and level of support at home, to name just a few. With **multiple correlation** (R), we determine the association between a variable that is of interest to us, in this case SAT score, and a combination of two or more predictor variables. Thus, with R we are usually able to more accurately reflect the true complexity of a situation than we would be if we limited ourselves to a Pearson r which involves only one predictor variable.

Multiple correlation (R) – The association between one criterion variable and a combination of two or more predictor variables.

With **partial correlation**, the second of our more complex correlational procedures, we can statistically remove the effect of a variable that is *not* of interest to us. For example, as a taxpayer it should be of interest to you to have some measure of the effectiveness with which your taxes are spent. If you live within a large school district there may well be a number of high schools. How can you compare the effectiveness of their academic programs? After all they may differ in many ways – size, age of buildings, years of teaching experience by the faculty, and so on. How can you possibly make a fair decision? One approach is to remove the effect of variables that are suspected of being important, but which are not of current interest. For instance, family income level is known to correlate with numerous variables associated with academic success of high school students. Not surprisingly, families with high incomes tend to provide more support at home in the form of computers, the parents tend to be more highly educated and are thus better able to assist their children academically, and these families even move to areas with new schools. Thus, it should come as no surprise that the most academically impressive high schools also tend to have students who come from the most affluent families. However, as a taxpayer you are probably not willing to simply conclude that those schools with the most affluent students are the most efficient users of your money. As a matter of fact, even though a school with students from less affluent families may not be achieving at quite the level as another with students from more affluent families, it may be excelling beyond what would be expected based upon the disparity in income. One way to determine whether this is the case is to employ a statistical procedure known as partial correlation. With this procedure we could remove the effect of family income to obtain a better view of which high schools were actually teaching most effectively.

Partial correlation – A procedure in which the effect of a variable that is not of interest is removed.

Regression

"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories instead of theories to suit facts."

A statistically significant Pearson correlation indicates that the variables are associated. In other words, if there is a significant correlation, knowing the value of one variable will assist in predicting the value of the other. However, a correlation does not indicate how this prediction is to be made. It was noted previously that in order to actually predict from one variable to another we use a procedure known as regression.

We will limit our discussion to the situation where there are two interval or ratio variables. It is assumed, therefore, that you have already calculated a Pearson r and that it was found to be significantly different from 0. A statistically significant Pearson r indicates that in the populations there is likely to be a relationship between the two variables. More specifically, the Pearson r indicates the extent to which there is a linear relationship between the two variables. Recall that if the value of one variable increases as the other increases, this is called a positive relationship (Figure 14.6). And if the value of one variable decreases as the other increases, this is called a negative relationship (Figure 14.7).

Figure 14.6 Example of a Positive Linear Relationship Between X and Y

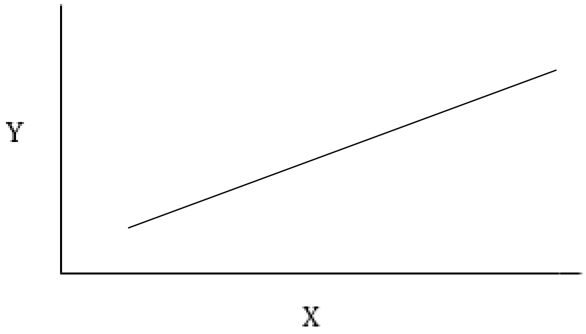
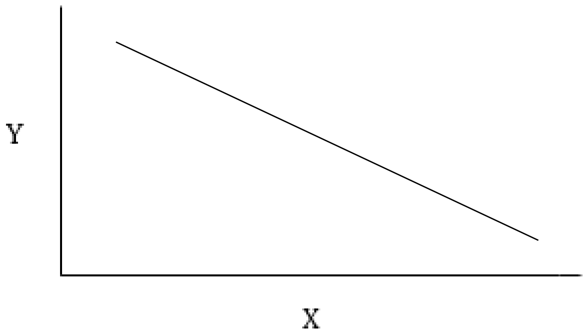
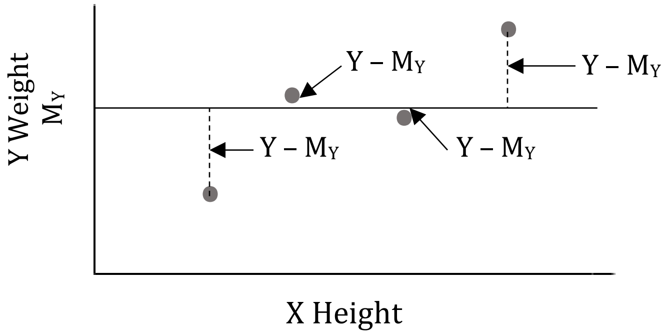


Figure 14.7 Example of a Negative Linear Relationship Between X and Y



Knowing that two variables are linearly related enhances our ability to predict from one to the other. To illustrate this, let us assume that we are *not* aware that there is an association between height and weight. In this case, regardless of a person's height, our best estimate of their weight would be the mean weight. Thus, regardless of a person's height (X), we would always predict the mean weight (M_Y) (Figure 14.8). Of course, for tall people we would tend to underestimate their weights, and for short people we would tend to overestimate their weights. Each error in estimation in Figure 14.8 is a deviation from the mean weight, or Y – M_Y. The sum of the errors would be $\sum(Y - M_Y)$. This term is not useful, of course, because it will always be equal to 0. However, you have learned that the sum of the squared deviations from the mean, $\sum(Y - M_Y)^2$ forms the basis for calculating the standard deviation and variance. In other words, the standard deviation and variance of the Y scores can be thought of as measures of our error of prediction assuming we always choose the mean weight (M_Y) regardless of the subject's height.

Figure 14.8 Example of Errors (Y – M_Y) Due to Predicting the Mean of Y (M_Y) Regardless of the Value of X

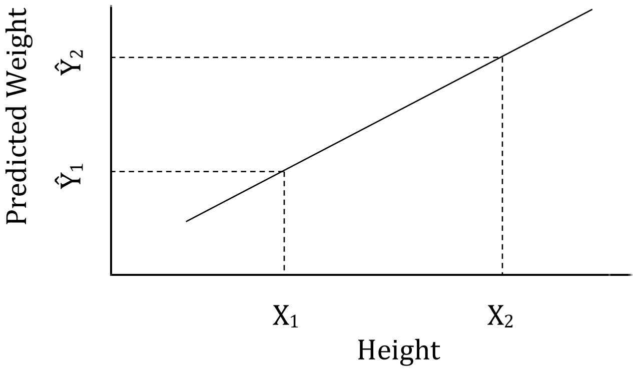


However, if there is a significant correlation between height (X) and weight (Y), then we know that a person's height (X) can assist in predicting that person's weight (Y). Thus, in this situation we know there is a better option than always choosing the mean weight (M_Y) regardless of the subject's height (X). Instead, with a positive correlation, as the height (X) increases so should our estimate of the weight (Y). **Simple linear regression** is the procedure used to derive an equation that enables us to predict from X to Y with optimal accuracy rather than choosing M_Y regardless of the value of X. Specifically, with simple linear regression one variable (X) is being used to predict the value of another variable (Y). Thus, for any value of X we will be able to use our equation to derive a predicted value of Y, for which we use the symbol $\hat{Y}$ (called Y hat). As we are dealing with linear regression, all of these $\hat{Y}$ values will fall along a straight line. This line is called the **regression line**. An example is illustrated in Figure 14.9. This example of a regression line rises to the right. This indicates that as the height (X) increases, so does the prediction of the weight ($\hat{Y}$).

Specifically, for a height of X_1 the predicted weight is $\hat{Y}_1$ and for a height of X_2 the predicted weight is $\hat{Y}_2$. Remember, the regression line consists of predicted values of Y.

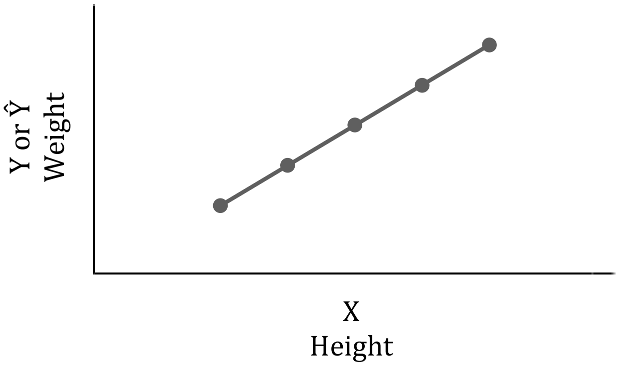
> *<u>Simple linear regression</u> – Procedure used to determine the equation for the regression line.*
>
> *<u>Regression line</u> – With simple linear regression, a straight line indicating the value of Y that is predicted to occur for each value of X. The symbol for the predicted value of Y is $\hat{Y}$.*

Figure 14.9 An Example of Using a Regression Line to Predict Y From X



In Figure 14.10, all of the actual subject weights fall along the regression line. In this situation, there would be no error in predicting from X to the Y values. In other words, if we know the subject's height, we can predict the subject's weight without any error. This would only be the situation if the correlation had a value of +1 (or –1). Of course, this rarely occurs in the real world. Instead, while in general taller people weigh more than those who are shorter, there are also individuals who are tall, but who are relatively light, and individuals who are short, but relatively heavy. In this more realistic situation, the Pearson correlation will have a value between –1 and +1.

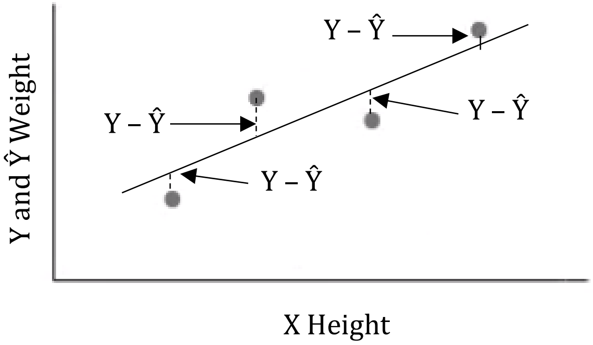
Figure 14.10 An Example of a Regression Line Permitting the Prediction of Y From X Without Any Error



Whenever the correlation is not + or –1 the observed data points do not all fall along the regression line (Figure 14.11). Since the regression line consists of the predicted value of Y for each value of X, any deviation from the line indicates that there was an error of prediction. Thus, the regression line consists of a series of Ŷ values, one for each value of X, and each deviation of an actual score (Y) from the predicted value (Ŷ) is an error of prediction. In other words, when Y – Ŷ does not equal zero, there is an error of prediction. As we are just as likely to underestimate as overestimate Y values, the ∑(Y – Ŷ) will be 0. However, the sum of the squared deviations from the predicted values of Y, which is written $\sum(Y - \hat{Y})^2$, can form the basis for calculating new measures of the standard deviation and variance. These can be thought of as measures of our error of prediction when using the regression line. In order to prevent confusion, the standard deviation for the error of prediction when using a regression line is called the **standard error of estimate** ($\sigma_{\hat{Y}}$).

<u>Standard error of estimate</u> ($\sigma_{\hat{Y}}$) – The standard deviation of Y scores around the regression line.

Figure 14.11 An Example of a Regression Line In Which There is Error (Y – Ŷ) in the Prediction of Y From X



The accuracy of our predictions will depend, first, upon how closely the observed data fall along a straight line and, second, how successful we are in defining the equation for the line that best fits our data. How tightly the data fall along a straight line is an empirical question and is out of our control. However, in those cases in which the Pearson correlation is large, either + or –, we know the data tend to fall tightly along a straight line and the accuracy of our predictions will be high. In contrast, in those cases in which the Pearson correlation is small the data do not fall as close to a straight line and the accuracy of our predictions will be lower. Regardless, we need an agreed-upon method for defining the equation for the line that best fits our data. In statistics, this regression line is defined as the straight line for which the sum of the squared errors of prediction, $\sum(Y - \hat{Y})^2$, is a minimum.

The calculation of maximums and minimums requires the use of calculus. Fortunately, we do not need to actually derive the equations for finding a straight line such that $\sum(Y - \hat{Y})^2$ is a minimum. Instead, we will simply make use of the equations that have been derived by others who used calculus. However, before doing so it is important that you understand the advantage of using a regression line.

Figure 14.8 is based upon the same hypothetical data points used in Figure 14.11. However, instead of showing the errors of prediction from the regression line (Figure 14.11), Figure 14.8 used the mean, M_Y, as the predicted weight for each subject regardless of the subject's height. It is evident that the total error of prediction when using M_Y (Figure 14.8) is greater than the total error of prediction when using the regression line (Figure 14.11). This will always be the case when the correlation is statistically significant.

Put another way, when the correlation is statistically significant the standard deviation based upon the deviations of Y scores from M_Y (which is an estimate of the population standard deviation, σ_Y) will always be greater than the standard error of estimate (the standard deviation based upon the deviations of Y scores from the regression line) (which is an estimate of the population standard error, $\sigma_{\hat{Y}}$). Thus, the error of prediction when using M_Y, which is σ_Y, will be greater than the error of prediction when using the regression line, which is $\sigma_{\hat{Y}}$, whenever the correlation is statistically significant. In other words, σ_Y {which is estimated by $\sqrt{[\sum(Y - M_Y)^2 / (n - 1)]}$} will be greater than $\sigma_{\hat{Y}}$ {which is estimated by $\sqrt{[\sum(Y - \hat{Y})^2 / (n - 1)]}$} whenever the Pearson r is statistically significant. This is evident from the following equation which defines the relationship of $\sigma_{\hat{Y}}$ and σ_Y:

$$\sigma_{\hat{Y}} = \sigma_Y \sqrt{(1 - r^2)}$$

So long as r does *not* equal 0, r^2 will be greater than 0, and $\sigma_{\hat{Y}}$ will be less than σ_Y.

For instance, if r is equal to –.5, then r^2 is equal to .25 and the equation becomes:

$$\begin{aligned}\sigma_{\hat{Y}} &= \sigma_Y \sqrt{(1 - .25)}\\ &= \sigma_Y \sqrt{(.75)}\\ &= \sigma_Y (.87)\end{aligned}$$

In other words, $\sigma_{\hat{Y}}$ is equal to 87% of σ_Y.

And what if r is equal to + or –1? Then r^2 is also equal to 1, and the equation becomes:

$$\begin{aligned}\sigma_{\hat{Y}} &= \sigma_Y \sqrt{(1 - 1)}\\ &= \sigma_Y \sqrt{(0)}\\ &= \sigma_Y (0)\\ &= 0\end{aligned}$$

This indicates that when r is equal to + or –1 all of the Y scores fall on the regression line and thus there is no error when predicting from X to Y, and $\sigma_{\hat{Y}}$ will equal 0.

The same relationship is true, of course, for the variances. Thus, the estimate of the population variance that is calculated using M_Y (i.e. σ_Y^2) will always be greater than what is called the **error variance** ($\sigma_{\hat{Y}}^2$), which is calculated using the regression line, so long as r does *not* equal 0.

Error variance ($\sigma_{\hat{Y}}^2$) – The variance of Y scores around the regression line.

Further, it can be shown that:

$$r^2 = \frac{\sigma_Y^2 - \sigma_{\hat{Y}}^2}{\sigma_Y^2}$$

An example will clarify the meaning of this equation. If all of the Y values fall directly along the regression line then r is equal to + or –1 and the error variance, $\sigma_{\hat{Y}}^2$, is equal to 0. This is the situation illustrated in Figure 14.10, and in this case there would be no error in prediction. If you know the value of X, you can predict the value of Y without any error. Specifically, in this case:

$$r^2 = \frac{\sigma_Y^2 - 0}{\sigma_Y^2}$$

$$= \frac{\sigma_Y^2}{\sigma_Y^2}$$

$$= 1$$

It was pointed out previously in this text that r^2 indicates the proportion of variability explained. With regression, we would say that r^2 indicates the proportion of the variability that has been accounted for, or eliminated, by using $\hat{Y}$ (the regression line) as our prediction rather than M_Y. As was just shown, when the correlation, r, is equal to + or –1, we can predict perfectly from X to Y. In other words, for each X value, the corresponding Y value is equal to $\hat{Y}$ and thus all of the variability has been explained. Stated differently, $\sigma_{\hat{Y}}^2$ would equal 0 and in this case r^2 is equal to 1. Furthermore, whenever $\sigma_{\hat{Y}}^2$ is small relative to σ_Y^2, it indicates that the predictions using the regression line are considerably *more* accurate than predictions using M_Y, and r^2 is therefore large (close to 1). However, whenever $\sigma_{\hat{Y}}^2$ approaches the size of σ_Y^2 it indicates that the predictions using the regression line are only marginally better than predictions using M_Y and, as a result, r^2 is small (close to 0).

To this point the discussion has been quite theoretical. This section began by explaining that you use linear regression following the determination that a Pearson r is statistically significant. With linear regression we are able to predict the value of Y that corresponds to a value of X. More specifically, it was noted that with a significant Pearson r the predictions based upon linear regression ($\hat{Y}$) are more accurate than if we simply chose the mean of the Y scores (M_Y)

regardless of the value of X. In other words, the standard error of estimate ($\sigma_{\hat{Y}}$) will be smaller than the standard deviation of Y scores from their mean (σ_Y). Finally, the relationship between linear regression and r^2 was reviewed.

It may have occurred to you that regression differs in an important way from correlation. With a Pearson r there are two variables, X and Y that are treated similarly since in a correlational study there is no independent or dependent variable. If the correlation is statistically significant, this indicates that the two variables are related, but this does not imply that one variable is causing a change in the other. With the subsequent linear regression the two variables are no longer treated similarly. Instead, we are using one variable to predict the value of the other. Thus one variable is the **predictor variable** (X) and the other is the **criterion** or **dependent variable** (Y). However, since regression is linked to a correlational study we still cannot conclude that a change in the predictor variable X is actually causing a change in the dependent variable Y.

Predictor variable (X) in regression – The variable (X) that is used to predict the value of the dependent or criterion variable (Y).

Criterion variable (Y) in regression – The variable (Y) whose value is being predicted by the predictor variable (X).

Dependent variable (Y) in regression – Another name for the criterion variable.

We will now conclude with a discussion of how to determine the actual equation for the regression line. It is important to note that if the Pearson r is not statistically significant then there is not sufficient evidence that a linear relationship exists between the variables, and thus there would be no point in identifying a regression equation.

Progress Check

1. If we have no idea what the relationship is between two variables, X and Y, then for every value of X, our best estimate of Y would be to choose _____.
2. If the Pearson correlation is statistically significant then using the _____ will lead to more accurate predictions than always choosing _____.
3. If the Pearson correlation is equal to + or –1, then the standard error of estimate will equal _____.

Answers: 1. the mean of Y 2. regression line; the mean of Y 3. zero

The Determination Of The Regression Equation

With a significant Pearson r we know that there is a relationship between the X and Y variables. However, the correlation does not tell us the equation of the straight line that best represents this relationship. In order to predict a value of Y, we need to use regression to define the precise relationship between the X and Y variables using the equation for a straight line, which is:

$$Y = bX + a$$

This equation indicates that the value of Y can be determined once the magnitude of X is given and two characteristics of the line are known. These characteristics are the **slope of the line**, 'b', and the Y intercept, 'a'. The slope of the line is defined as the ratio of how much the Y variable changes as the X variable changes. It is also called the **regression weight**:

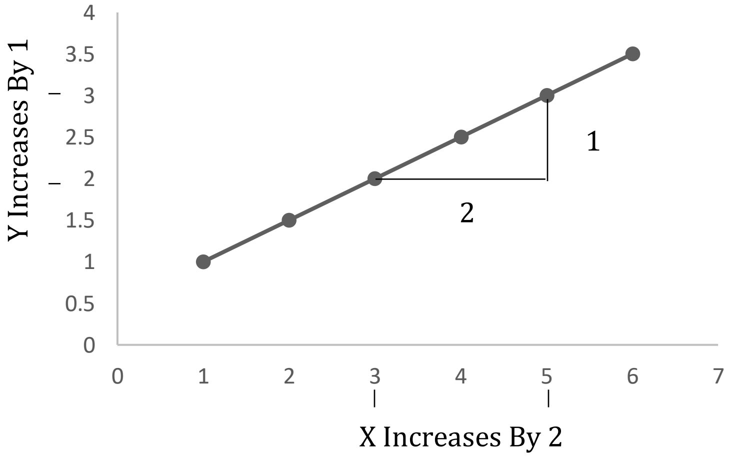
$$b = \frac{\text{Change in Y}}{\text{Change in X}}$$

Thus, if Y increases by 1 when X increases by 2, the line has a slope of ½ or 0.5. Similarly, if Y increases by 3 when X increases by 6, the slope is also 0.5. In each case the ratio of the change in Y divided by the change in X, which equals 'b', remains 0.5. This is shown in Figure 14.12.

Slope of the line – One of the two determinants of the equation for a straight line. It is the ratio of the change in the Y variable divided by the change in the X variable. It has the symbol 'b' in the equation Y = bX + a. It is also called the regression weight.

Regression weight – Another term for the slope of the regression line.

Figure 14.12 Example of a Line with a Slope of ½ or 0.5



The second determinant of the equation for a straight line, the **Y intercept**, is the value of Y when X is equal to 0. In other words, it is the value of Y when the line crosses the Y axis. If you extend the line in Figure 14.12, you will see that the Y intercept, 'a', is equal to 0.5.

Y intercept – One of the two determinants of the equation for a straight line. It is the value of Y when X is equal to 0. It is, therefore, the value of Y when the *line crosses the Y axis. It has the symbol 'a' in the equation Y = bX + a.*

Determining The Slope Of The Regression Line

Regression describes the procedure for finding the equation for the straight line that best fits our data. As was noted previously, the best-fitting regression line is defined as the straight line for which the sum of the squared errors of prediction, $\sum(Y - \hat{Y})^2$, is a minimum. It was also pointed out that while the calculation of maxima (the plural of maximum) and minima (the plural of minimum) requires the use of calculus, we do not need to perform the derivations of the equations for finding a straight line with $\sum(Y - \hat{Y})^2$ as a minimum. Instead, we can use the equations that have been found by others using calculus. Specifically, the slope of the regression line is:

$$b = r\left(\frac{\sigma_Y}{\sigma_X}\right)$$

If 'b' is positive, the regression line will rise or slope upward to the right as in Figure 14.6. If 'b' is negative, the regression line will slope downward to the right as in Figure 14.7.

Of course, we usually do not know the population standard deviations (σ_Y and σ_X). However, we can estimate these population standard deviations from the sample data using s_Y and s_X. For instance, we previously calculated a statistically significant Pearson r of .79 for hypothetical quiz and exam scores for seven students taking statistics (Tables 14.2 – 14.5). And in order to calculate the Pearson r we also calculated values for s_Y and s_X. We will now proceed to find the regression line for these data.

The equation for 'b', the slope of the regression line becomes:

$$b = r\left(\frac{s_Y}{s_X}\right)$$

Substituting the estimates of the standard deviations derived from the sample data (calculations associated with Tables 14.4 and 14.5):

$$b = .79\left(\frac{8.99}{1.35}\right)$$

$$= .79\ (6.66)$$

$$= 5.26$$

Determining The Y Intercept Of The Regression Line

The equation for 'a', the Y intercept, is:

$$a = M_Y - bM_X$$

The values for the mean of the exam scores (M_Y) and the mean of the quiz scores (M_X) come from Table 14.3. We just calculated the value for 'b'. The equation for 'a' thus becomes:

$$a = 84.29 - (5.26)\,(8.14)$$

$$= 84.29 - 42.82$$

$$= 41.47$$

The Regression Equation

As was noted previously, the general equation for the regression line is $\hat{Y} = bX + a$. Substituting for 'b' and 'a', which were just calculated, we have:

$$\hat{Y} = 5.26\,X + 41.47$$

Based upon this equation, a quiz score of 0 would be associated with an exam score of 41.47. And with each increase of 1 point on the quiz, we predict an increase of 5.26 points on the exam. This regression line is graphed in Figure 14.13. (However, it is important to note that actual predictions should be limited to the range of quiz grades used to calculate the original correlation (Table 14.2).

Figure 14.13 The Regression Line for Hypothetical Quiz and Exam Grades

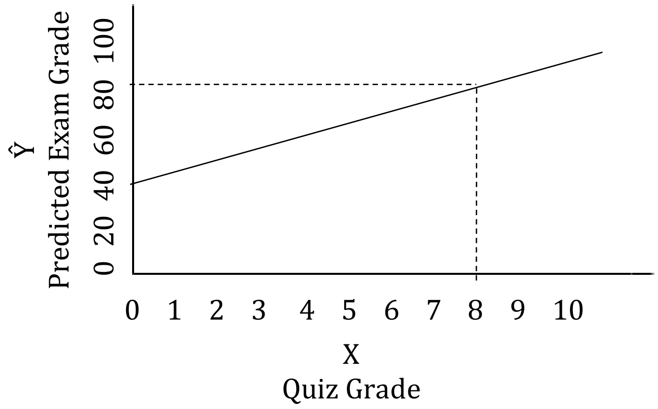


Figure 14.13 shows that as the quiz grade increases, so does the predicted exam grade. The relationship between quiz grades and exam grades is also evident from the value of the Pearson r, which is .79, but that relationship is now presented graphically. The regression line in Figure 14.13 can be used to obtain a quick estimate of a student's exam grade. For instance, inspection of Figure 14.13 indicates that a student with quiz grade of 8 is predicted to obtain an exam grade in the 80s. We can find a more precise prediction by using the regression equation we just determined:

$$\hat{Y} = 5.26\,X + 41.47$$

For a quiz grade of 8, we have:

$$\hat{Y} = 5.26\ (8) + 41.47$$

$$= 42.08 + 41.47$$

$$= 83.55$$

The predicted value of 83.55 obtained from the regression equation thus confirms our visual estimation using Figure 14.13 but is, of course, more precise.

Determination Of The Standard Error Of Estimate

If you refer back to Table 14.2, you will find that two students had quiz scores of 8. However, neither of these students obtained the predicted exam grade of 83 or 84. Remember, it is only in the case where the Pearson r is equal to + or −1 that you would expect the actual data points to fall exactly along the regression line. Our correlation of .79 is quite large. In other words there is a good fit between the data points and the regression line. Nevertheless, because the correlation is not 1 we do not have a perfect match between the predicted and actual data.

As you will recall, the calculated regression equation results in the best-fitting line, and thus there will be less error of prediction using this than any other line. And as was shown previously, the error of prediction when using the regression line, which is called the standard error of estimate ($\sigma_{\hat{Y}}$), will be less than if the mean of Y was always chosen as the estimate, which would lead to an error equal to the standard deviation (σ_Y).

It is important to recognize that the accuracy of our predictions when using the regression equation is limited by the strength of the original correlation, r. This can be illustrated with the equation for the standard error of estimate ($\sigma_{\hat{Y}}$), which was provided previously:

$$\sigma_{\hat{Y}} = \sigma_Y \sqrt{}(1 - r^2)$$

When using sample data, s_Y provides an estimate of σ_Y. The equation, therefore, becomes:

$$\sigma_{\hat{Y}} = s_Y \sqrt{}(1 - r^2)$$

With our current example, s_Y was found to be 8.99. And r was .79. Therefore:

$$\sigma_{\hat{Y}} = 8.99 \sqrt{}[1 - (.79)^2]$$

$$= 8.99 \sqrt{}[1 - .62]$$

$$= 8.99 \sqrt{}.38$$

$$= 8.99\ (.62)$$

$$= 5.57$$

As was previously noted, whenever the Pearson r is statistically significant the standard error of estimate ($\sigma_{\hat{Y}}$), which is a measure of how well we can predict using the regression equation, will be less than the standard deviation (σ_Y), which is a measure of how well we can predict if the mean of Y is chosen for every value of X. In the present example this is confirmed, for based upon our sample data the standard error of estimate is 5.57, which is less than the standard deviation of 8.99.

Finally, once again it is important to note that in our example of the quiz and exam scores, the range of the original quiz scores was from 6 to 10. You should restrict your prediction of exam grades to those future students who have quiz scores between 6 and 10. In other words, even though Figure 14.13 extends from a quiz score of 0 up to a quiz score of 10, and the regression equation we derived will calculate a predicted exam score corresponding to any quiz score, we have no knowledge of what the relationship would be beyond our original range of quiz scores (6 to 10). You should, therefore, limit any predictions to this range of values.

Reporting The Results Of A Pearson r Followed By Regression

When reporting the results of a correlational study we would state whether the correlation was significant and, if so, note the value of the coefficient of determination and that a linear regression was then performed. The regression equation would be followed by providing the standard error of estimate. Specifically, for our example of quiz (X) and exam (Y) scores (Table 14.2), we would state that the calculated Pearson r was found to be significant ($r(5) = .79$, $p < .05$, $r^2 = .62$). The regression equation was $\hat{Y} = 5.26\ X + 41.47$, and the standard error of estimate was 5.57. If we were going to publish our findings, then we should use a statistical package to gain greater precision and to provide a precise p-value. With SPSS we would again find that the Pearson r was significant ($r(5) = .80$, $p = .033$, $r^2 = .63$) and the regression equation would now be $\hat{Y} = 5.32\ X + 41.00$. The standard error of estimate is now 5.98. Most of these values correspond closely to what we calculated. However, the value for the standard error of estimate differs substantially from what we previously found. This discrepancy is explained in the later section of this chapter that describes how to conduct regression with SPSS.

Multiple Correlation and Regression

The present chapter began with a discussion of correlation, the extent to which two variables are related. And it was noted that the amount of variability accounted for will often be enhanced by including additional variables. For instance, knowing only the student’s SAT exam score accounts for some variability in college achievement. However, the amount of variability accounted for can be increased by including other variables such as high school grade point average (GPA) and a measure of the difficulty of the courses that were taken. This is an example of multiple correlation – determining the degree of association between one variable and a number of other variables.

Just as you can have multiple correlation, you can also have **multiple linear regression**. With multiple linear regression, which is covered in more detail in the appendices, the equation describing the linear relationship between two or more X variables and a single Y variable is determined. The advantage of multiple linear regression over simple linear regression is that the inclusion of additional variables will often lead to a more accurate prediction of the value of Y. The disadvantage of multiple linear regression is that the calculations are more complex than the procedures that have been reviewed in this chapter. You are strongly encouraged, therefore, to use a computerized statistical package rather than hand computation when using multiple regression.

Multiple linear regression – A procedure in which several variables (Xs) are used to predict the value of another variable (Y).

Purpose And Limitations Of Using Simple Linear Regression

1. *Provides an equation so that the value of Y can be predicted.* The Pearson correlation provides a measure of the strength and direction of an association between two interval or ratio variables. Simple liner regression provides an equation for this association.
2. *Not a measure of cause and effect.* Simple linear regression follows the finding of a statistically significant Pearson r. Due to a lack of control in a correlational design a researcher is not justified in coming to a cause-and-effect conclusion concerning the variables. The regression equation allows the prediction of Y from X but does not indicate that X is causing Y.
3. *Prediction is limited to the range of the original values.* The regression equation should not be used for values of X that are beyond the range of the data that were used in the calculation of the Pearson r.

Assumptions Of Simple Linear Regression

1. *Interval or ratio data.* The data are on an interval or a ratio scale of measurement.
2. *Data are paired.* The data come as pairs, usually two measures on the same individual.
3. *Linear relationship.* The Pearson correlation and linear regression assume that the two variables are linearly related.
4. *Significant Pearson r.* Simple linear regression is only used if the Pearson r has been found to be statistically significant.

Conclusion Of Regression

The focus of this section has been upon linear regression. When there is a statistically significant linear relationship between the X and Y variables, then it is possible to have a better prediction of Y than always choosing the mean of Y. Simple linear regression is utilized to determine the actual equation relating X and Y so we can make these more accurate predictions of Y.

In addition, multiple linear regression, a procedure in which two or more X variables are used to predict the value of Y, was briefly discussed.

Glossary Of Terms

Coefficient of determination – The square of the correlation. It indicates the proportion of variability in one variable that is explained or accounted for by the variability in the other variable.

Coefficient of nondetermination – The proportion of the variability of one variable not explained or accounted for by the variability of the other variable. For the Pearson r, it is equal to 1 – r^2 *.*

Correlation – A measure of the degree of association among variables. A correlation indicates whether a variable changes in a predicable manner as another variable changes.

Correlation coefficient – A single number that indicates the degree to which two variables are related.

Covariance – A statistical measure indicating the extent to which two variables vary together.

Covary – If knowledge of how one variable changes assists you in predicting the value of another variable, the two variables are said to covary.

Criterion variable (Y) in regression – The variable (Y) whose value is being predicted by the predictor variable (X).

Dependent variable (Y) in regression – Another name for the criterion variable.

Error variance ($\sigma_{\hat{Y}}^2$) *– The variance of Y scores around the regression line.*

Multiple correlation (R) – The association between one criterion variable and a combination of two or more predictor variables.

Multiple linear regression – A procedure in which several variables (Xs) are used to predict the value of another variable (Y).

Negative correlation – A relationship between two variables in which as one variable increases in value, the other variable decreases in value. Also, as one variable decreases in value, the other increases in value.

Partial correlation – A procedure in which the effect of a variable that is not of interest is removed.

Positive correlation – A relationship between two variables in which as one variable increases in value, so does the other variable. Also, as one variable decreases in value, so does the other.

Predictor variable (X) in regression – The variable (X) that is used to predict the value of the dependent or criterion variable (Y).

Regression – Procedure researchers use to develop an equation that permits the prediction of one variable of a correlation if the value of the other variable is known.

Regression line – With linear regression, a straight line indicating the value of Y that is predicted to occur for each value of X. The symbol for the predicted value of Y is $\hat{Y}$.

Regression weight – Another term for the slope of the regression line.

Restriction of the range – Reducing the range of values for a variable will reduce the size of the correlation.

Rho (ρ) – Symbol used for the population correlation.

Simple linear regression – Procedure used to determine the equation for the regression line.

Slope of the line – One of the two determinants of the equation for a straight line. It is the ratio of the change in the Y variable divided by the change in the X variable. It has the symbol 'b' in the equation $Y = bX + a$.

Standard error of estimate ($\sigma_{\hat{Y}}$) – The standard deviation of Y scores around the regression line.

Y intercept – One of the two determinants of the equation for a straight line. It is the value of Y when X is equal to 0. It is, therefore, the value of Y when the line crosses the Y axis. It has the symbol 'a' in the equation $Y = bX + a$.

Questions – Chapter 14 – Correlation

(Answers are provided in Appendix J.)

1. When knowledge of the outcome of one event assists in predicting the outcome of another event, then we say _____.
   a. The two events are causally related
   b. The two events are correlated
   c. The two events are independent
   d. The two events are meaningful

2. In order to use the Pearson r the data must be either _____.
   a. nominal or ordinal
   b. ordinal or interval
   c. interval or ratio
   d. nominal or ratio

3. The magnitude of the Pearson r indicates the _____ between X and Y.
   a. Durability of the relationship
   b. Direction of the relationship
   c. Degree to which there is a non–linear relationship
   d. Degree to which there is a linear relationship

4. The magnitude of the square of the Pearson r indicates the _____.
   a. Percent of the variance in Y explained by the variance in X

b. Degree to which X and Y are affected by a third variable, Z
c. Degree to which the experimenter has utilized appropriate experimental design
d. Extent to which error occurred in the study

5. If knowing how one variable changes aids us in predicting how another variable will change, we say that the two variables _____.
a. Are causally linked
b. Covary
c. Are identical
d. Should be merged into one variable

6. If you are interested in whether two variables are correlated and if both variables consist of ordinal data you use _____.
a. ANOVA
b. Spearman r
c. Pearson r
d. Chi-square

7. If you are interested in whether two variables are correlated and if you are dealing with two interval or ratio variables you would employ _____.
a. ANOVA
b. Spearman r
c. Pearson r
d. Chi-square

8. A correlation of _____ indicates that there is not any association between the two variables.
a. 0
b. 1
c. 2
d. 3
e. 4

9. A correlation of _____ indicates that there is a perfect association among the variables.
a. 0
b. 1
c. 2
d. 3
e. 4

10. In a _____ correlation, as one variable increases, so does the other.
a. Neutral
b. Negative
c. Positive
d. Strong

11. In general, the more flaws a diamond has, the lower its value. This is an example of a _____ correlation.
a. Neutral
b. Negative
c. Positive
d. Strong

12. With a (an) _____ the researchers know that a prediction can be made and how well it can be made, but they do not know what the actual prediction would be. In order to make a prediction, we employ the statistical procedure called _____.
a. Inferential statistic; correlation
b. Descriptive statistic; correlation
c. Regression; correlation
d. Correlation; regression

13. If we conclude, based upon our samples, that a correlation exists when in fact there is no such correlation, we have made a _____.
a. Type I error
b. Type II error
c. Type III error
d. A correct decision

14. With a (an) _____ design, we are not asking if the distributions that we have observed are different. Instead, we are asking if the variables are related or associated.
a. Experimental
b. Descriptive statistical
c. Correlational
d. ANOVA

15. The square of a correlation is called the _____.
a. Spearman correction
b. coefficient of determination
c. co–variance
d. regression equation

16. The _____ measures what proportion of variance in one variable is explained or accounted for by the other variable.
a. Spearman correction
b. coefficient of determination
c. co–variance
d. regression equation

17. The removal of extreme scores usually reduces the size of a correlation. This is called _____ .
a. The compression effect
b. Range limitation
c. Deviation control
d. Restriction of the range

18. A group of hypothetical students were asked their high school GPAs and their most recent statistics quiz score:

GPA	Quiz Score
4.0	10
3.75	9
3.5	9
3.25	7
3.0	8
2.5	5

What is the Pearson correlation for these two variables?
a. .93
b. .95

c. –.92
d. .03

Note: It is important to save your work for problem 18 as it will be needed for subsequent questions dealing with regression.

Questions – Chapter 14 – Regression

(Answers are provided in Appendix J)

19. The greater the magnitude of the correlation, ignoring the sign, between X and Y, the _____.
a. Farther the data points are from the regression line
b. Closer the data points are to the regression line
c. More closely the data points around the regression line look like a circle
d. The lower the ability to predict from X to Y

20. To actually predict from one variable to another, we use a procedure known as _____.
a. Regression
b. Correlation
c. Dependency analysis
d. Post hoc testing

21. The sum of the errors, $\sum (Y - M_Y)$, will always be equal to _____.
a. 0
b. 1
c. 2
d. 3
e. None of the above

22. If you don't have any other information, your best prediction of Y would be _____ for every value of X.
a. Mean of X
b. Mean of Y
c. Mean of X + Y
d. Mean of Y^2

23. The standard deviation for the error of prediction when using a regression line is called the _____.
a. standard deviation of estimate
b. standard deviation of error
c. standard error of estimate
d. standard variation of estimate

24. When the correlation is 1, all of the observed data points fall along the _____.
a. X axis
b. Y axis
c. horizontal line for the mean of Y
d. regression line

25. The regression line is defined as the straight line for which the sum of the squared errors of prediction, $\sum (Y - \hat{Y})^2$, is a _____.

a. Minimum
b. Maximum
c. Mean of the minimum and maximum
d. None of the above

26. Whenever the Pearson r is significant, $\sqrt{[\sum (Y - M_Y)^2 / n]}$ _____.
a. will be greater than $\sqrt{[\sum (Y - \hat{Y})^2 / n]}$
b. will equal $\sqrt{[\sum (Y - \hat{Y})^2 / n]}$
c. will be less than $\sqrt{[\sum (Y - \hat{Y})^2 / n]}$

27. When r^2 is equal to 1, the standard error of estimate ($\sigma_{\hat{Y}}$) is equal to _____.
a. 0
b. 1
c. 2
d. 3

28. We are more accurate in making predictions when r^2 is _____.
a. Small
b. Of intermediate size
c. Large
d. It depends upon the specific question being asked.

29. In the general equation for a straight line, $Y = bX + a$, the slope is indicated by _____.
a. 'b'
b. 'X'
c. 'a'
d. None of the above

30. The procedure for finding the equation for the straight line that best fits our data is called _____.
a. Correlation
b. Linear maximization
c. Finding the Y intercept
d. Regression

31. If the correlation between X and Y is zero, then for any value of X the best prediction for the value of Y would be the _____.
a. standard error of estimate
b. minimum value of Y
c. maximum value of Y
d. mean value of Y

32. In problem #18 that involved GPA and quiz grades, what is the value of the constant in the regression line?
a. 3.09
b. –0.08
c. –2.29
d. 43.15

33. Problem #18 involved GPA and quiz grades. What is the value of the slope of the regression line for these data?
a. 3.09
b. –0.08

c. −2.29

d. 43.15

Problems 34 – 42 utilize SPSS.

Using SPSS With The Pearson Correlation

To Begin SPSS

Step 1 Activate the program, close the central window, and click on the **Variable View** option at the bottom left of the window.

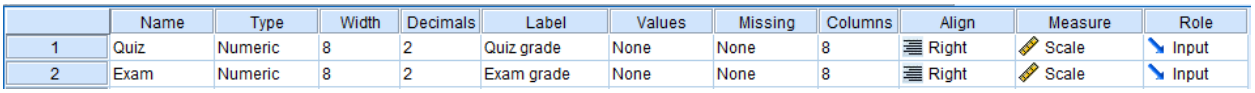
Step 2 Click on the first empty cell under the column heading ‘Name’. You now type the name of the first variable for which you have data. We are going to utilize the same data and labels as were previously employed in Table 14.2. These data dealt with whether there is a correlation between quiz and exam grades for students in a statistics class. We are calling these variables ‘Quiz’ and ‘Exam’. Therefore, type ‘Quiz’ in the first empty cell under ‘Name’.

Step 3 Click on the first empty ‘cell’ under the column heading ‘Label’. In this cell you can type a more extensive description of your variable. In our case, type ‘Quiz grade’.

Step 4 Click on the first empty ‘cell’ under the column heading ‘Measure’. As we are dealing with ratio data for the quiz grades, select ‘Scale’ as is shown in Figure 14.14.

Step 5 Repeat Steps 2 – 4 except that you type ‘Exam’ in the first empty cell under ‘Name’ and ‘Exam grade’ for the label. As before, select ‘Scale’ in the column under the column heading ‘Measure’ as we have ratio data for the exam grades. The result is shown in Figure 14.14.

Figure 14.14 The Variable View Window



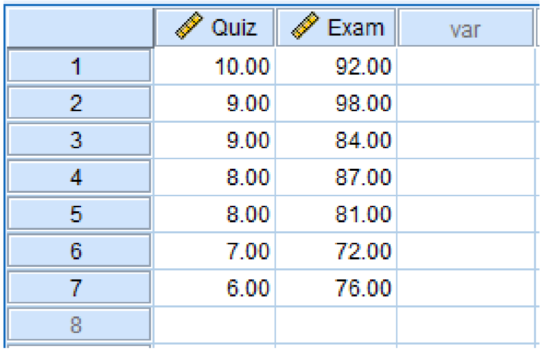
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Quiz	Numeric	8	2	Quiz grade	None	None	8	Right	Scale	Input
2	Exam	Numeric	8	2	Exam grade	None	None	8	Right	Scale	Input

To Enter Data In SPSS

Step 6 Click on the ‘Data View’ option at the lower left corner of the window. The variables ‘Quiz’ and ‘Exam’ will be present.

Step 7 For each of the seven subjects in the study type their quiz and exam grades in the appropriate columns (Figure 14.15).

Figure 14.15 Entering Data



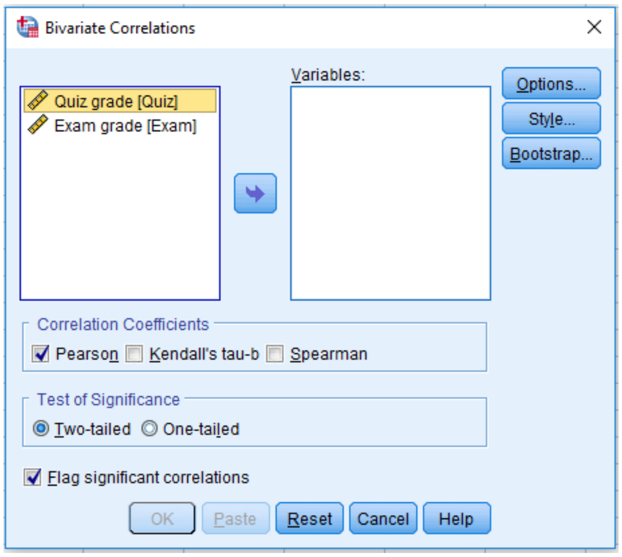
	Quiz	Exam	var
1	10.00	92.00	
2	9.00	98.00	
3	9.00	84.00	
4	8.00	87.00	
5	8.00	81.00	
6	7.00	72.00	
7	6.00	76.00	
8			

To Conduct A Pearson Correlation

Step 8 Click the cursor on '**Analyze**' along the row of SPSS commands above the data you entered, then move to '**Correlate**', then click on '**Bivariate**' as we are dealing with two variables, the quiz and exam grades.

Step 9 A new window will appear (Figure 14.16). On the left side of this window is a list of all of the variables that have been entered into SPSS. In order to conduct a Pearson r we must indicate to SPSS which variables we wish to examine. As we only have two variables this is accomplished by moving our two variables to the empty box on the right side of the window. To do so check that our first variable, 'Quiz Grade', is highlighted and then click on the central arrow (Figure 14.17). We then move 'Exam Grade' to the right side box in the same manner. The result will be that each label will move to the appropriate box on the right–hand side of the window, as is shown in Figure 14.18. Check to be sure that the appropriate options are indicated, in our case '**Pearson**', '**Two-tailed**' and '**Flag significant correlations**'. Then click '**OK**' which is located at the bottom of the window.

Figure 14.16 The Bivariate Correlation Window



Bivariate Correlations

Quiz grade [Quiz]
Exam grade [Exam]

Variables:

Correlation Coefficients

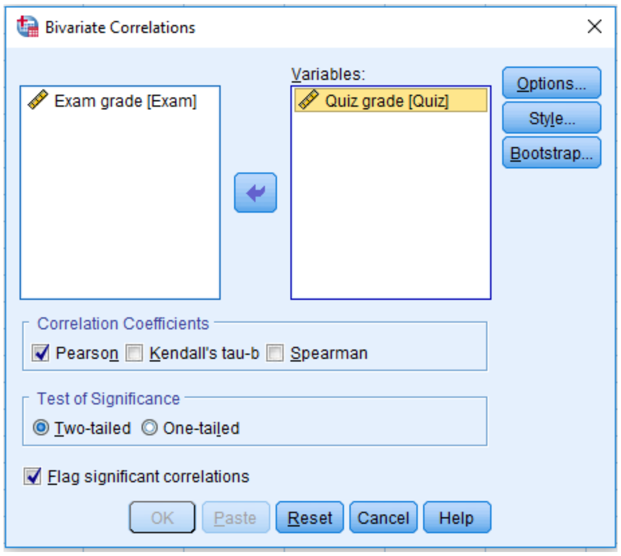
☑ Pearson ☐ Kendall's tau-b ☐ Spearman

Test of Significance

● Two-tailed ○ One-tailed

☑ Flag significant correlations

Figure 14.17 The Bivariate Correlation Window, Continued


Bivariate Correlations

Exam grade [Exam]

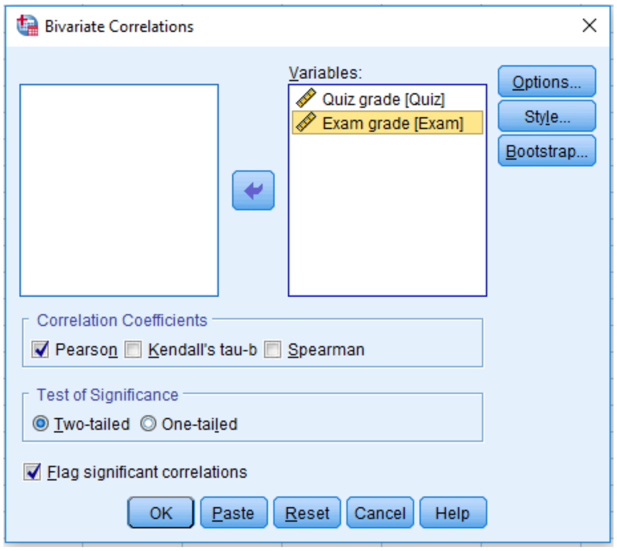
Variables:
Quiz grade [Quiz]

Correlation Coefficients
☑ Pearson ☐ Kendall's tau-b ☐ Spearman

Test of Significance
● Two-tailed ○ One-tailed

☑ Flag significant correlations

Figure 14.18 The Completed Bivariate Correlation Window


Bivariate Correlations

Variables:
Quiz grade [Quiz]
Exam grade [Exam]

Correlation Coefficients
☑ Pearson ☐ Kendall's tau-b ☐ Spearman

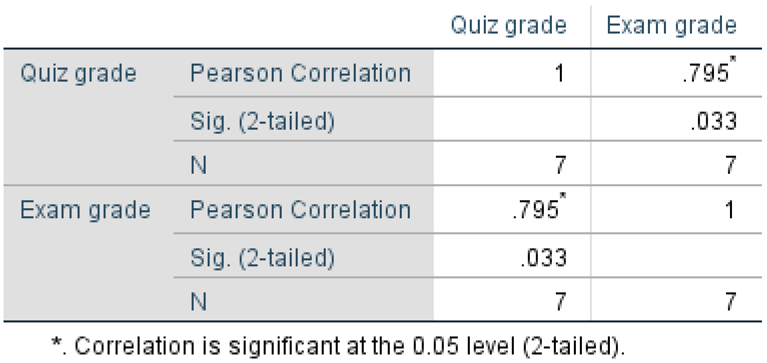
Test of Significance
● Two-tailed ○ One-tailed

☑ Flag significant correlations

Step 10 SPSS calculates the desired Pearson r as shown in Table 14.10. The table takes some practice to get used to. In the left-most column, find 'Quiz grade'. Directly to the right is printed 'Pearson Correlation'. Continuing in the same row there is a number 1, which indicates that quiz grades are perfectly correlated with quiz grades. This is obvious and is not of interest. Continuing in the same row is the number .795*. This indicates that the Pearson r between quiz grades and exam grades is .795 and the * indicates that this correlation is significant at the .05 level.

The next row indicates that the p-value of a correlation of .795 is .033, which is less than an alpha of .05. The final row in the Quiz grade section is the number of pairs of scores, in this case 7. The same information is then presented again in three rows that begin with ‘Exam grade’. Clearly, if the correlation between quiz and exam grades is .795, then the correlation between exam and quiz grades is also .795. You should verify that the outcome using SPSS is essentially the same as we found previously.

Table 14.10 SPSS Output; Pearson Correlation



		Quiz grade	Exam grade
Quiz grade	Pearson Correlation	1	.795*
	Sig. (2-tailed)		.033
	N	7	7
Exam grade	Pearson Correlation	.795*	1
	Sig. (2-tailed)	.033	
	N	7	7

*. Correlation is significant at the 0.05 level (2-tailed).

Note: Save this SPSS data file. It will be used in the following SPSS section dealing with regression.

SPSS Problems – Correlation

A magazine recently listed the horsepower and mileage of sports cars equipped with a turbo and a manual transmission:

Horsepower	MPG
200	27
265	24
172	33
227	25
197	27
305	21

34. What is the correlation between horsepower and miles per gallon (MPG)?
    a. .932
    b. –.950
    c. –.916
    d. .025

35. Is the correlation statistically significant with alpha equal to .05?
    a. yes
    b. no

Note: Save this SPSS data file. It will be used in the following SPSS section dealing with regression.

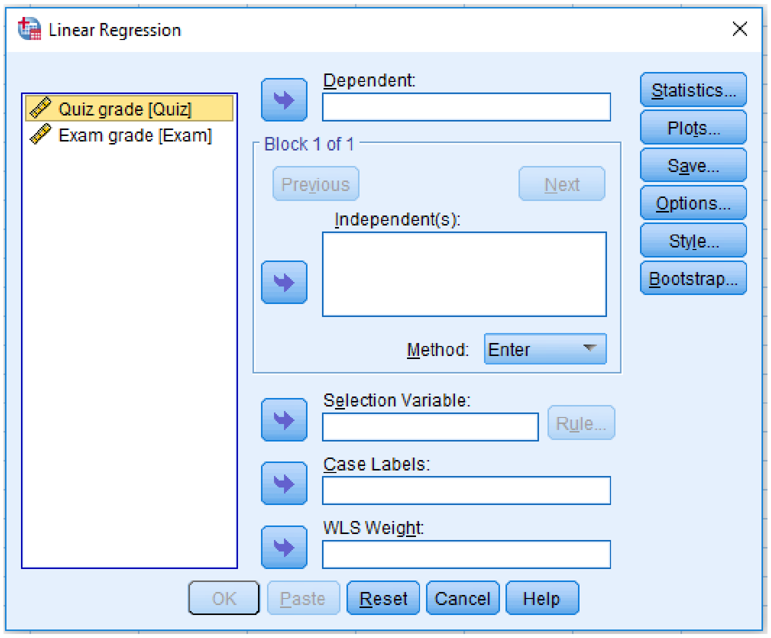
Using SPSS For Linear Regression

We will continue to use the data from Table 14.2 to illustrate how SPSS can be used to calculate a linear regression.

Step 1 Retrieve the SPSS data file that was previously created for the data in Table 14.2 (If you did not save this file, you will need to go back to the SPSS section dealing with correlation and follow the steps to enter the data.)

Step 2 Click on ‘**Analyze**’, then on ‘**Regression**’ and finally on ‘**Linear**’. A new window appears (Figure 14.19).

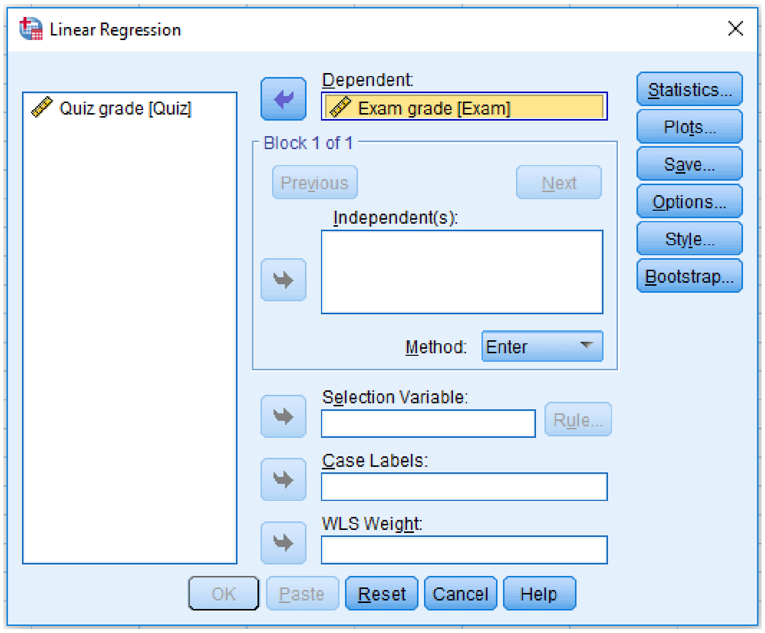
Figure 14.19 The Linear Regression Window



Method: Enter

Step 3 As we are trying to predict exam grades from quiz grades, highlight ‘Exam grade’ and click on the top arrow. ‘Exam grade’ will move to the box under ‘Dependent’ (Figure 14.20).

Figure 14.20 The Linear Regression Window – Continued

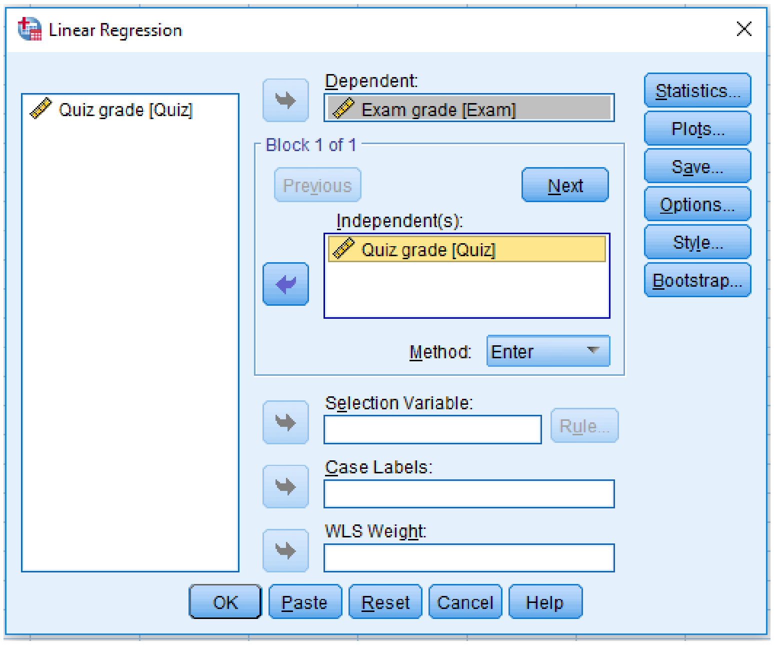


Dependent: Exam grade [Exam]

Method: Enter

Step 4 Highlight ‘Quiz grade’ and click on the second arrow. ‘Quiz grade’ will move to the box under ‘Independent(s)’ as it is our predictor variable (Figure 14.21).

Figure 14.21 The Linear Regression Window – Completed



Dependent: Exam grade [Exam]

Independent(s): Quiz grade [Quiz]

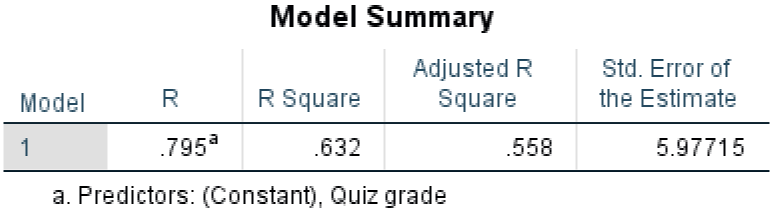
Method: Enter

Step 5 Click on ‘**OK**’ and the SPSS linear regression analysis will appear. We are only interested in the last three of the four sections of the output. Table 14.11 indicates the value of the correlation for the linear regression is .795. (SPSS uses the symbol R to signify this correlation.) This is the same value we found for the correlation previously with SPSS and, except for rounding error, with our hand calculations. In addition, Table 14.11 shows that the variability in the exam scores that can be accounted for by the differences in the quiz scores (R Square) was .632. This closely matches our calculated value of r^2 which was .62. Furthermore, a value for the 'Adjusted R Square' is given. Finally, the standard error of estimate is provided. Approximately 68% of the estimated exam grades will fall within plus or minus one standard error of estimate, in this case 5.97715, and approximately 95% of the exam grades will fall within plus or minus two standard errors of estimate. (You may have noticed that the value of the standard error of estimate in Table 14.11, which is 5.97715, differs substantially from the value of 5.57 that we calculated. This is due to SPSS using the 'Adjusted R Square' of .558 to calculate the standard error of estimate while we utilized an r^2 of .62. You are encouraged to recalculate the standard error of estimate using the value from Table 14.11 to confirm that this is the case.)

Table 14.12 provides a statistical test (ANOVA) of whether using the regression line provides a better estimate of the exam grades than if the researcher always chose the mean value of the exam grades regardless of the value of the quiz grades. As the significance level (p-value) is reported to be .033, which is less than .05, the conclusion based upon the ANOVA is that the regression equation provides a better estimate.

Table 14.13 provides the actual regression coefficients. In this chapter we have learned that the equation for the regression line is provided by the equation $\hat{Y} = bX + a$. The values for 'a' and 'b' are listed in the column with the heading 'B'. Specifically, the value for the constant 'a' is 41.000 and the value for the slope 'b' is 5.316. Thus the regression equation becomes $\hat{Y} = 5.316X + 41.000$. And note that these values for 'a' and 'b' are similar to the values that we previously found except for slight discrepancies due to rounding error when the calculations were completed by hand. We do not need to be concerned with the remainder of Table 14.13.

Table 14.11 SPSS Output; R, R Square, and Standard Error of Estimate

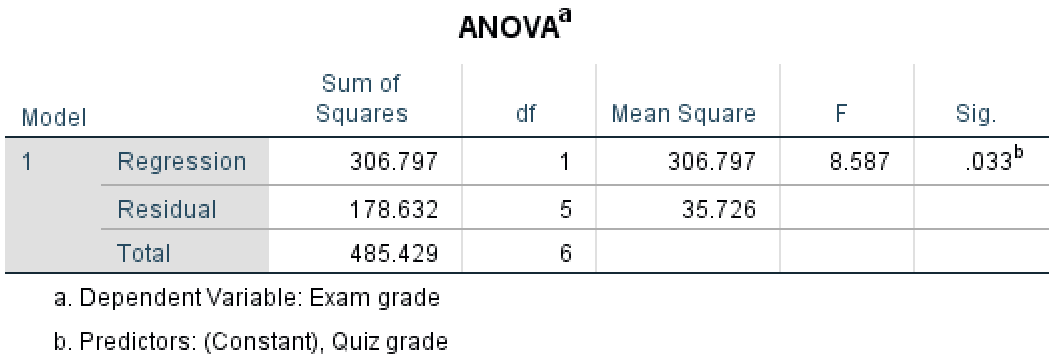


Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.795[a]	.632	.558	5.97715

a. Predictors: (Constant), Quiz grade

Table 14.12 SPSS Output; Test of Significance



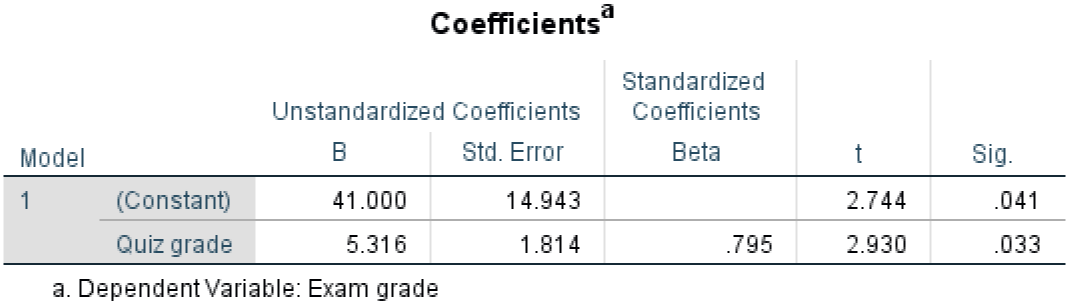
ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	306.797	1	306.797	8.587	.033[b]
	Residual	178.632	5	35.726		
	Total	485.429	6			

a. Dependent Variable: Exam grade

b. Predictors: (Constant), Quiz grade

Table 14.13 SPSS Output; Regression Coefficients



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	41.000	14.943		2.744	.041
	Quiz grade	5.316	1.814	.795	2.930	.033

a. Dependent Variable: Exam grade

Caution: For SPSS, which variable you identify as the DV and which as the IV in regression is critical as the values for 'a' and 'b' will be affected.

Step 6 Exit SPSS. There is no need to save the output or data.

SPSS Problems – Regression

For the following problems, utilize the data previously entered for questions 34 – 35 that dealt with the association between horsepower and miles per gallon. Use SPSS for problems 36 and 37. Subsequent problems may need calculations to be completed by hand.

36. What is the value of the Y intercept in the regression line?
    a. 3.086
    b. −0.075
    c. −2.286
    d. 43.152

37. What is the value of the slope of the regression line?
    a. 3.086
    b. −0.075
    c. −2.286
    d. 43.152

38. What is the predicted miles per gallon for a car with 300 horsepower?
  a. 20.65
  b. 22.39
  c. 23.45
  d. 24.60

39. What is the predicted miles per gallon for a car with 400 horsepower?
  a. 16.92
  b. 13.15
  c. A value should not be calculated as 400 horsepower is beyond the range of the original data
  d. Less than 10

40. Finally, calculate the predicted miles per gallon for a car with 200 horsepower.
  a. 27.33
  b. 28.15
  c. 29.25
  d. 30.67

CONCLUSION

Chapter 15 – Congratulations, the Big Picture and Next Steps: Recapitulation and Final Considerations

Chapter 15
Congratulations,
The Big Picture and Next Steps:
Recapitulation and Final Considerations

"The science of statistics is the chief instrumentality through which
the progress of civilization is now measured, and by which
its development hereafter will be largely controlled."
S. N. D. North

General Review

You are to be congratulated! This is the final chapter of a demanding book. It is the author's hope that you have not only mastered the techniques that have been presented but that you have also gained an appreciation of the usefulness of statistical analyses of data.

As you are quite aware there are numerous statistical procedures, each appropriate for a different situation. An overview table (Appendix L) was utilized to assist you in seeing how the various procedures are related and so that you would be better able to understand the context within which they are used. At the broadest level there are two types of statistics, descriptive and inferential. As you learned, descriptive statistics consist of those procedures that are used to summarize a set of data. Measures of central tendency, such as the median and mean, as well as measures of variability, including the range and standard deviation, are examples of descriptive statistics. Most of the text, however, was devoted to a review of inferential statistics. These are the procedures used with experimental, quasi-experimental and most correlational designs. Inferential statistical procedures enable us to conclude whether a relationship observed in a sample is likely to generalize to a population. Examples of these procedures include the Pearson correlation, the chi-square test of independence and the ANOVAs.

Of course, as the overview table (Appendix L) indicates, the specific descriptive or inferential procedure that is appropriate will also depend, in part, upon the type of data that have been collected. Statistical procedures deal with data measured at the nominal, ordinal, interval or ratio levels. The nature of the research question that is being examined will usually determine a specific level of measurement. Once the question is identified there is often little choice in the level of measurement that will be employed. However, since the amount of information conveyed by the data differs with the level of measurement, the power of the statistical tests that are matched with the various levels of measurement will also differ (reviewed in Appendix E). Thus the power of the statistical tests that are used with interval and ratio data is greater than the power of the tests appropriate for ordinal or nominal data. As a result, fewer subjects will be needed when collecting interval or ratio data compared to ordinal or especially nominal data. For instance, a two-way between-subjects ANOVA is more powerful, and thus more efficient, than a chi-square test of independence.

Regardless of the inferential statistical procedures used, their strength is in enabling us to predict. Properly employed statistical procedures enable us to generalize findings from our sample to a population and, accordingly, to predict the likelihood of future occurrences. This has proven to be incredibly valuable. With the assistance of statistics we are able to predict the academic success of college applicants who have not yet finished high school, we can predict the effectiveness of medical treatment options, and we can predict economic outcomes, to name just a few uses. In other words, statistics are immensely practical. Because of this, they are also ubiquitous. Since you cannot hide from them, a better approach is to learn about statistics so that you can benefit from their potential. Hopefully this book has assisted you in achieving this goal.

"If you think that statistics has nothing to say about what you do or how you could do it better, then you are either wrong or in need of a more interesting job."

Stephen J. Senn

Future Directions

The author had a number of purposes in writing this book. A major goal, of course, was to provide an introduction to the most essential statistical procedures. This book should have provided you with the background needed to understand much of the research in your specific field of interest. The author also aspired to provide an introduction to statistics organized in such a way that the relationships between different statistical procedures would be evident (refer to Appendixes L and M). This perspective should provide a good foundation for those of you who plan to learn more about statistics in the future. In fact, if this text has served to enhance your interest in this field and, as a result, you are excited about continuing to explore the field of statistics, or research methodology, then I am most gratified. If, on the other hand, you see this as your last formal exposure to statistics I hope that you have gained an appreciation of the usefulness of statistical analyses and the knowledge that you have mastered a number of the field's important concepts. However, regardless of whether this is your last, or just your first, course in statistics you should be aware that due to time limitations there are numerous statistical procedures that could not be included.

At the nominal level of measurement a very useful procedure is the **Fisher exact test**. It is an alternative to the 2 X 2 chi-square test of independence and is usually employed with studies which have small data sets.

Fisher exact test – *An alternative to the 2 X 2 chi square test of independence that is used when there is a particularly small data set.*

There are also numerous additional procedures for use with interval or ratio data that you may see in the literature. For instance, **factor analysis** permits the identification of which variables, out of a set of predictor variables, statistically group together. Each related group of variables is called a **factor**. (The term 'factor' is being defined differently here than when we discussed factorial ANOVAs.) This technique was developed by Spearman and Burt in the 1930s to try to ascertain whether intelligence consisted of a series of largely independent characteristics or whether there was some shared component underlying the more specific attributes. Factor analysis involves highly complex calculations and is, therefore, reliant upon computer-assisted data analysis.

Factor analysis – *Statistical procedure that groups the initial variables into a smaller set of underlying variables called factors.*

Factor - *one of a smaller number of underlying variables derived from analysis of the larger set of initial variables.*

In addition, you are likely to see larger ANOVAs than were covered in this text. We discussed one- and two-way designs. Three-way and even larger ANOVAs are encountered in the literature. There is no theoretical limit to the number of independent variables that can be included in an ANOVA, but the interpretation of the interactions quickly becomes problematical. Also, we did not have time to cover what is called the **mixed ANOVA**. This is a very useful, factorial design in which there are both between-subjects and within-subjects factors.

Mixed ANOVA – *Factorial ANOVA in which there are both between-subjects and within-subjects factors.*

Just as we learned that ANOVAs can be expanded so that the effects of two or more independent variables can be simultaneously analyzed, a technique known as **MANOVA** permits the simultaneous analysis of more than one dependent variable. This can be essential to analyzing some research designs, but once again the calculation are best left to a computer.

MANOVA – *An extension of ANOVA in which there is more than one dependent variable.*

This by no means exhausts the additional statistical options. It may be helpful to think of each statistical procedure as a tool. You have now acquired a basic, general tool kit, such as initial buyers of a home or car often purchase. You have the equivalent of a hammer, a few pliers and a couple of screwdrivers. The more you learn in your field of interest, the more statistical 'tools' you are likely to acquire. Many are quite specialized. But, as with the tools of a master mechanic, each has its purpose. The more techniques you learn, the more flexible you become in analyzing data.

Overview: This Statistics Book Makes Use Of Mathematics But Is Not A Mathematics Book

It may have occurred to you that while this statistics book uses a great deal of mathematics it is different from other mathematics texts you have studied. That is because this is not truly a mathematics book. By this it is meant that this statistics book has a different orientation than one commonly written by mathematicians. Mathematicians generally seek universal solutions through logical analysis. In this text we have not developed any such general solutions. We have not, therefore, been functioning as mathematicians. Instead, we have been the beneficiaries of their efforts. Fundamentally, in this text we have learned how to employ the solutions developed by mathematicians. In order to employ the solutions correctly we have had to recognize what type of research problem we were facing, but what I have not attempted to do is to provide an in-depth explanation of the underlying logic of any of the statistical procedures.

This distinction between a mathematician's analysis of a statistical problem and our use of statistical procedures is evident at a number of levels. Most significantly, mathematicians find 'truth' through a method called **deduction**.

<u>Deduction</u> – A method of thinking in which conclusions are logically derived from general statements that are assumed to be true.

Box Dealing With Contributions Of The Greeks

Deduction was developed by the Greeks and is the fundamental method used by mathematicians. This method emphasizes human reason, or **rationalism.** Before the Greeks, the Egyptian and Babylonian civilizations had found solutions for mathematical operations and had used basic algebra. They also had determined how to find areas and volumes, and thus had a value for π that was accurate enough for their purposes. For the Egyptians, this was 3.16 while for the Babylonians it was simply 3, the same value as is found in the Bible. Thus, for the Egyptians and Babylonians, mathematics was simply a tool. They used mathematics practically, for measurement, finance and astronomy. Their mathematics were useful, but limited.

<u>Rationalism</u> – A method for finding truth that emphasizes logical thinking rather than observation.

While the Greeks acknowledged their debt to other civilizations, especially to the Egyptians, they proceeded to revolutionize mathematics and many other fields with their emphasis upon human reasoning. In their view the senses are inadequate to find truth and, instead, we should emphasize the human capacity for logic. The immediate consequence was that the Greek civilization created western philosophy, political analysis as well as our conception of mathematics. A mathematician begins with some basic assumptions, called axioms, and from them develops theorems. So long as the initial axioms are correct, deductions based upon logical thinking will lead to general solutions that we can be completely confident are true. Thus, a mathematician seeks definite knowledge through a rational process.

The results of this emphasis upon reason were impressive. By 300 BCE, Greeks such as Thales and Pythagoras had made significant advances, especially in geometry. This effort culminated with Euclid's *Elements* in which 467 theorems are deduced from an initial set of 10 axioms. Throughout the Middle Ages, the Renaissance and for several subsequent centuries Euclid's *Elements* served as the foundation for a Western education in logical thinking. It is undoubtedly one of the most significant and influential books ever written.

The emphasis of Greek thinking was not upon practical gain. It was, instead, focused upon the acquisition of pure knowledge. Nevertheless, the Greek emphasis upon human reason, and particularly their reliance upon deduction in mathematics, has had an enormous practical as well as theoretical legacy. For instance, it is a commonly repeated myth that Columbus, in order to get support for his proposal to reach China by sailing west, had to first convince the king and queen of Spain that the earth was round rather than flat. The truth is that it was generally accepted in fifteenth century Europe by those who were educated that the earth was essentially a sphere and, in fact, there were several estimates of its size. The most commonly accepted estimate of the earth's circumference came from Ptolemy (90 – 168 CE), a Greek scholar who had resided in Alexandria, Egypt. (This is the same Ptolemy who is known for his geocentric model of the universe.) However, an even earlier estimate of the earth's circumference had been made by Eratosthenes (276? –195? BCE), a Greek scholar who had also lived in Alexandria, Egypt (reviewed in Boorstin, 1983).

Eratosthenes, like most educated Greeks, accepted that the earth was essentially a sphere. He learned, additionally, that on the summer solstice the sun's rays reached all the way to the bottom of a deep well at Syene (modern Aswan), which was a known distance (approximately) to the south of Alexandria. He realized that this meant that the sun had to be positioned directly over the well on that day (Figure 15.1). He also understood that this information, along with his training in geometry, would permit him to estimate the earth's circumference. Specifically, on the day of the summer solstice he measured the length of the shadow made by an obelisk (some references indicate a vertical rod) of known height at Alexandria. From these measurements he determined that the angle between the obelisk and the sun's rays was slightly greater than 7° (Figure 15.2). The rest was geometry. We will call the angle between the obelisk and its shadow 'a'. We will round this value to be 7°. Therefore, we also know that another angle 'A' has the same value (Figure 15.3). Hence, the angle between the line extending from the obelisk in Alexandria to the center of the earth, and the line from the well at Syene to the center of the earth is also 7°. Thus, the distance that Syene is to the south of Alexandria corresponds to 7 / 360 or approximately 1 / 50 of the circumference of the earth since there are 360° in a circle. Eratosthenes noted that the circumference of the earth is, therefore, 50 times the distance that Syene is to the south of Alexandria. We now know that the circumference of the earth is about 40,075 km (approximately 24,900 miles). The accuracy of Eratosthenes' estimate is still being debated since there is not a universally agreed upon conversion between his unit of measurement of distance (the stade) and the units of measurement we currently use. The most generally accepted outcome is an estimate of 28,700 miles, and thus Eratosthenes' error was only about 15%. (His error may have been substantially less.) Regardless, I think you will agree that this is a remarkable achievement considering it was accomplished over 2000 years ago using only a few simple measurements. Of course, Eratosthenes also benefited from a long history of mathematical progress that was a direct result of the Greek emphasis upon deduction.

Figure 15.1 The Position of the Sun on the Summer Solstice

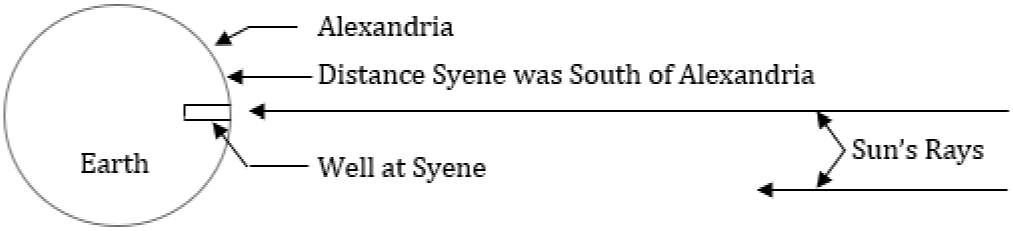


Figure 15.2 Diagram of the Analysis Employed by Eratosthenes

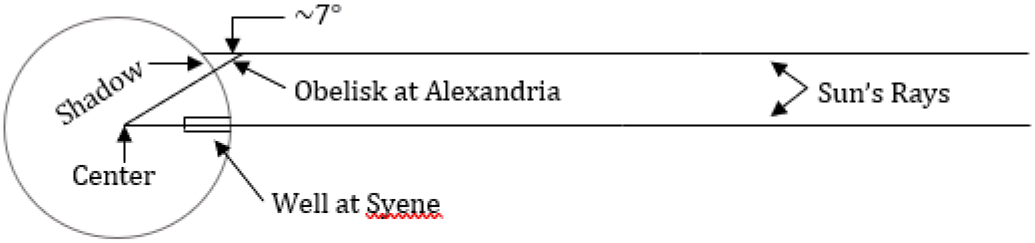
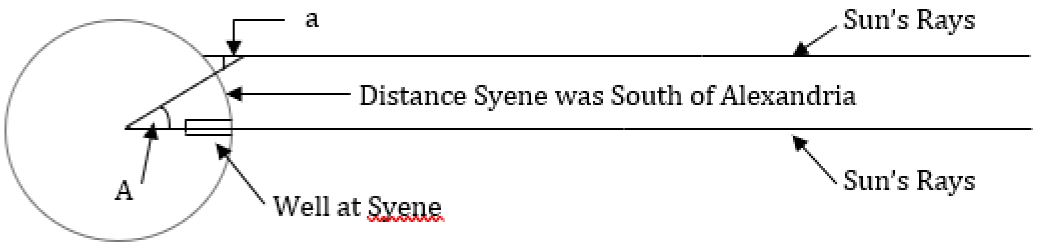


Figure 15.3 Eratosthenes' Geometric Analysis



And what about Columbus? Fortunately for him, in fifteenth century Europe Ptolemy was the accepted authority in geography and his estimate of the earth's circumference was only 18,000 miles. This was an error of over 27%. In addition, Ptolemy thought Asia was larger than it actually is. Together, these errors had the effect of dramatically shrinking the distance between Eastern Asia and Western Europe. And Columbus went further and argued that even Ptolemy's underestimate of the size of the earth was too large!

A Portuguese commission rejected Columbus' proposal to sail west to reach the Indies. The royalty of Spain were also skeptical, but Columbus finally convinced them that the distance from Spain to Japan and China was much smaller than it actually is, and thus that a voyage from Spain west to Japan and China was feasible. As a glance at a globe will indicate, Columbus and the crews of his three small, leaky ships were very fortunate indeed that two unknown continents lay between Spain and Japan or China. Otherwise Columbus and his sailors would probably have never been heard of again.

In summary, Greek rationalism led to several estimates of the size of the earth. The estimate made by Eratosthenes was surprisingly accurate. However, it was Ptolemy's estimate that was generally accepted in fifteenth century Europe. If the true size of the earth had been known, then it is unlikely that Columbus' plan to sail west to China would have been supported. Put another way, this 'discovery' of the New World was, in part, due to several errors. Fortunately for Columbus, some people are just lucky!

We, as users of statistics, function quite differently than mathematicians or the ancient Greeks for we emphasize **induction**. With inferential statistics we begin with a limited sample of data and attempt to generalize the outcome to some population. Thus, instead of beginning with axioms that are assumed to be true, we start with hypotheses and then use observations and statistical analyses to determine their likelihood. We understand that the outcome of this process is not the definite knowledge that a mathematician seeks but rather a probabilistic statement. While the statistical procedures developed by the mathematician will always be correct so long as the initial assumptions are met, the specific outcomes we obtain by using the statistical procedures reviewed in this text will only have a probability of leading to a correct conclusion. For example, if the Overview Table indicates that an ANOVA is the appropriate test we can be confident that the equations used are correct because they are based on mathematicians' deductive efforts. However, the outcome of using an ANOVA will be in terms of a probability. We have learned, for instance, that with α equal to .05, there is a 5% chance of making a Type I error. No matter how carefully the study is conducted, or the data analyzed, we will never achieve certainty. This is because unlike the mathematician, who approaches a problem from a rational perspective, as statisticians we approach our problems from an empirical perspective.

Induction – A method of thinking in which conclusions are derived from generalizations based upon limited statements or observations that are assumed to be true. Induction is fundamental to science, as observations are used to develop general laws of nature.

With **empiricism**, we gain knowledge through observation. Everyone is at some level an empiricist. You choose your friends based upon your observations of their behavior; perhaps you chose your car after reading reviews; and you will receive grades based upon your professors' observations of your learning. In this text you have been introduced to sophisticated methods of empirical inquiry. These are the correlational, quasi-experimental and experimental designs. Used correctly these procedures, paired with the appropriate statistical analyses, greatly enhance the likelihood of gaining knowledge through observation. Nevertheless, empiricism is always based upon limited observations and thus will never permit the absolute confidence that comes with the deductive method employed by mathematicians.

Empiricism – A method for finding truth that emphasizes the importance of observation.

Though statistical analysis does not lead to certainty, it does lead to a probabilistic understanding of situations that has revolutionized many fields of study. In fact, it is not an exaggeration to say that statistical thinking is largely responsible for the transformation of fields such as economics, sociology and psychology into sciences, and it has dramatically affected others such as anthropology, political science and history. This is an amazing outcome for an offshoot of mathematics that began with the analysis of games of chance.

Cautions In Using Statistics

"Statistics are no substitute for judgment."

Like any powerful tool, statistical analysis must be used intelligently. When using statistical procedures we are often interested in generalizing from a sample(s) to a population(s). It is essential to recognize that our confidence in doing so is dependent, in large part, upon the manner in which the sample(s) was chosen. A biased sample does not provide the necessary basis for confidently making such a generalization. Consequently, in this text we have repeatedly emphasized the importance of randomly selecting the subjects who will be included in the sample, a view emphasized by R. A. Fisher (1935). In fact, the statistical procedures you have learned assume that the samples have been randomly selected. Nevertheless, many studies, perhaps most, do not use random sampling. This is not because the researchers are ignorant of the need for random sampling. Instead, it is a consequence of the difficulty of obtaining random samples in a real-world setting. For instance, researchers at a college might want to generalize the results of their study to the entire population of Americans. To do so, they recognize that their sample should be a random selection of everyone residing in the United States. But, how could they practically collect such a sample? Yet if they don't, how justified are they in claiming that their findings will generalize?

In psychology many studies are conducted with samples drawn from students taking an introductory college-level course. To what population would it be appropriate to generalize the results? Not only are the subjects all in college and thus likely to be younger than the American population in general, they also are more likely to be female. Clearly, not only is this not a random sample, it is not even remotely representative of the entire American population. Instead it is what is often called a **sample of convenience**. The researchers are aware that they do not have a random sample, but what alternative do they realistically have? Texts such as this one often make it sound as if selecting the sample is straightforward. Actually, selecting an appropriate sample can be extremely challenging.

Sample of convenience – A sample that is chosen because it is easily available rather than because it is optimal.

At least researchers are aware of the problem of selecting an appropriate sample. They are, however, commonly not aware of their own biases. There are numerous examples of researchers finding what they were looking for. Some striking instances are reviewed in the *Mismeasure of Man* by Stephen J. Gould. In this book Gould explains how numerous, competent researchers of the 19th century published findings of skull volumes supporting the commonly held view that women and minorities were intellectually inferior to white males. As Gould is careful to point out, this was usually not the result of any conscious manipulation of the data. In other words, in the vast majority of cases there is no evidence that fraud was involved. Instead, the researchers apparently were convinced of what the findings would be before they collected the data and, not surprisingly, interpreted their findings according to these preconceived views. Biases such as these are, in some respects, more problematic than outright dishonesty. Hopefully, none of us is ever going to publish a fraudulent paper. However, any of us could unknowingly publish results or conclusions affected by biases that we are not even aware we have.

In order to control against such biases the researcher can conduct what is called a **'blind' study**. There are a number of variants, but the essential idea is that neither the subject nor the researcher knows to which group the subject is assigned when the data are collected. If properly used this technique will prevent biased observation. However, 'blind' studies are more difficult to conduct and, in some cases, are not feasible. For instance, if a study involves a comparison of men and women in face-to-face interactions, without careful precautions it is difficult to imagine that the individual collecting the data will not be aware of whether the subjects are men or women.

> *'Blind' study – A study in which the data are collected in such a way that the subject's assignment to the control or experimental condition is not known. There are several variations of 'blind' procedures. They are all employed to reduce bias.*

It is also important to recognize that while statistical procedures are valuable in making predictions, these predictions are much more accurate for groups than for individuals. For instance, it has been shown that exercise in the elderly is beneficial for a number of conditions. Thus, for two groups of the elderly who are equivalent except for how much they exercise, I can confidently predict that the group that exercises more will be healthier. Nevertheless, there are likely to be individuals who exercise and yet who are not healthy, and individuals who do not exercise and yet remain healthy. Therefore, while I can confidently predict which *group* will be healthier, it is much more challenging to predict how healthy any particular *individual* will be. Remember, most of the statistical procedures we reviewed compared *sample means, not individual scores*. A statistically significant outcome thus does not indicate that every individual in the sample had exactly the same reaction to an intervention, just that overall there is an effect of the intervention. As a consequence, following a statistically significant finding we can be confident that an independent variable had an effect. However, this does not indicate that the effect occurred, or occurred equally, for each individual.

And always remember that statistics deals with probabilities, not certainty, and recognize that this lack of certainty does not negate the usefulness of the procedures that you have reviewed in this text.

A Final Thought

"They do not with regard to the phenomena seek for their reasons and causes

but forcibly make the phenomena fit opinions and preconceived notions…"

Aristotle

We have just reviewed some very important philosophical issues – rationalism/empiricism, deduction/induction, and the danger of bias. What may not be clear is that this book was written from a particular perspective which is fundamental to the field of statistics. This perspective, or assumption, is so basic that it may seem obvious: use of statistical procedures presupposes that facts matter. Put another way, it has been assumed throughout this book that an emphasis upon data is critical to understanding the world, and people. This is, clearly, an empirical view. This view was discussed previously and, I suspect, the vast majority of readers accepted it uncritically. However, I believe that upon closer inspection you will agree that some people seem immune to facts or feedback – they seem to hold views that are simply not open to being changed no matter how much data are presented.

"People almost invariably arrive at their beliefs not on the basis of proof

but on the basis of what they find attractive."

Blaise Pascal

Clearly, this book assumes that facts should modify our opinions, and not the reverse. But once again I think you will agree that many, perhaps all, people sometimes seem to see what they want to see, and they ignore what is uncomfortable or threatening to their views and values. In fact, this occurs so commonly that in psychology it is given a name, the **confirmation bias**.

<u>Confirmation bias</u> – Selecting only evidence that supports, or confirms, one's pre-existing beliefs.

"His mind, in a sense, was too masterful – it imposed itself upon realities."

Richard Hofstadter describing John C. Calhoun

This book has presented a variety of tools to help you see the world more clearly and to improve the quality of your decisions. It is likely that you have spent much of your time learning the details of the many statistical procedures that were presented. This is common in a first exposure to statistics. At the same time, I have tried to also assist you in gaining a broader perspective by emphasizing an understanding of how these procedures are related. This is summarized in the Overview Table. What I am now suggesting is that this is fundamentally a book on critical thinking. And if you now question the basis for the decisions that others as well as you make, then mastering the material presented in this book will have been a particularly valuable experience.

Conclusion

"The most useful skill we could teach is the habit of asking oneself and others, how do you know? If knowledge comes from intuition or anecdote, it is likely wrong."

Sharon Begley

You have now completed this introduction to statistics. Statistical procedures are powerful tools in our quest to understand the world we live in. Used correctly, statistics can be extremely valuable. Though an incorrect conclusion is always possible, properly used statistical analysis will lessen the likelihood of making errors and will greatly enhance our ability to predict relationships among variables. Used incorrectly, statistics can be quite detrimental, as the 'cooked' books at failed companies such as ENRON indicate. Like any capability, your knowledge of statistics needs to be paired with integrity and judgment. Finally, in research it is important to keep focused upon the big picture. Quality studies require an insightful research idea, careful implementation of procedures and correct statistical analysis of the resulting data. And always remember,

"Not everything that can be counted counts, and not everything that counts can be counted."

George Gallup

Glossary Of Terms

'Blind' study – A study in which the data are collected in such a way that the subject's assignment to the control or experimental condition is not known. There are several variations of 'blind' procedures. They are all employed to reduce bias.

Confirmation bias – Selecting only evidence that supports, or confirms, one's pre-existing beliefs.

Deduction – A method of thinking in which conclusions are logically derived from general statements that are assumed to be true.

Empiricism – A method for finding truth that emphasizes the importance of observation.

Factor - one of a smaller number of underlying variables derived from analysis of the larger set of initial variables.

Factor analysis – Statistical procedure that groups the initial variables into a smaller set of underlying variables called factors.

Fisher exact test – An alternative to the 2 X 2 chi square test of independence that is used when *there is a particularly small data set.*

Induction – A method of thinking in which conclusions are derived from generalizations based upon limited statements or observations that are assumed to be true. Induction is fundamental to science, as observations are used to develop general laws of nature.

MANOVA – An extension of ANOVA in which there is more than one dependent variable.

Mixed ANOVA – Factorial ANOVA in which there are both between-subjects and within-subjects factors.

Rationalism – A method for finding truth that emphasizes logical thinking rather than observation.

Sample of convenience – A sample that is chosen because it is easily available rather than because it is optimal.

References

Fisher, R. A. (1971) (1935). The design of experiments (9th ed.). Macmilan.

Gould, S. J. (1981). *The mismeasure of man.* New York: Norton.

Questions – Chapter 15

(Answers are provided in Appendix J.)

1. Which of the following is the correct order for the power of tests, from those with the least to those with the most power?
   a. Ordinal; nominal; interval/ratio
   b. Interval/ratio; nominal; ordinal
   c. Nominal; interval/ratio; ordinal
   d. Nominal; ordinal; interval/ratio

2. _____ organizes the initial variables into statistically related underlying factors.
   a. Kruskal-Wallis H test
   b. ANOVA
   c. Spearman
   d. Factor analysis

3. A method of thinking in which conclusions are logically derived from general statements that are assumed to be true.
   a. Empiricism
   b. Deduction
   c. Induction
   d. None of the above

4. A method for finding truth that emphasizes logical thinking rather than observation.
   a. Empiricism
   b. Induction
   c. Rationalism
   d. None of the above

5. Euclid's *Elements* dealt with _____.
   a. Statistics
   b. Algebra
   c. Geometry
   d. Poetry

6. Instead of beginning with axioms that are assumed to be true, in statistics we begin with data that have been observed. This is the process of _____.
   a. Induction
   b. Deduction
   c. Rationalism
   d. Thought for which the ancient Greeks are famous

7. Statistical analysis does not lead to certainty, instead it leads to a (an) _____.
   a. Absolute truth
   b. Probabilistic understanding
   c. Inability to predict the future
   d. Unsubstantiated opinion

8. Scientists would prefer to have a _____sample, but they often must employ a _____ sample.
   a. Convenience; random
   b. Biased; unbiased
   c. Random; convenience
   d. None of the above

9. The essential idea of a (an) _____ study is that which group the subject is assigned to is not known by the researcher when the data are collected.
   a. Well–controlled
   b. Experimental
   c. Correlational
   d. Blind

10. George holds strong political views and only listens to politicians who have the same views he does. This is an example of _____.
   a. confirmation bias
   b. deduction
   c. induction
   d. a sample of convenience

11. I am interested in determining what students think of the new menu being offered for lunch so I ask my friends for their opinions. This would be an example of _____.
   a. confirmation bias
   b. induction
   c. a sample of convenience

APPENDIXES

Appendixes A – E, Additional Statistical Procedures

A. Kruskal-Wallis H Test: Analysis of a Difference Design with One Independent Variable, Two or More Samples, and Ordinal Data

B. Phi Correlation: Identifying the Strength of an Association when there are Nominal Data

C. Spearman Correlation: Identifying the Strength of an Association when there are Ordinal Data

D. Multiple Linear Regression

E. An Introduction to Power Analysis – Minimum Appropriate Sample Sizes

Appendixes F – H, Statistical Symbols, Equations and Measures of Effect Size

F. Statistical Symbols Used in this Book

G. Definitional Equations and, Where Appropriate, Their Computational Equation Equivalents

H. Inferential Statistical Procedures and Their Measures of Effect Size

Appendixes I – J, Glossary and Answers to Chapter/Appendix Problems

I. Glossary of Terms

J. Answers to Chapter/Appendix Problems

Appendixes K – L, Tables and Overview for Choosing the Correct Procedure

K. Statistical Tables

L. Overview Table

M. Comparison of ANOVAs

N. Choosing the Correct Inferential Procedure Table and Practice Choosing the Correct Procedure

Appendix A
Kruskal-Wallis H Test: Analysis of a Difference Design with One Independent Variable, Two or More Independent Samples, and Ordinal Data

"Nothing has such power to broaden the mind as the ability to investigate systematically and truly all that comes under thy observation in life."
Marcus Aurelius

Introduction

Our review of inferential statistics began with the procedures used with nominal (frequency) data. We then turned to the procedures used with interval or ratio data. These data are commonly called scores. In this appendix we review a procedure used with ordinal data. Ordinal data consist of ranks rather than scores. Researchers utilize ordinal data in two situations. In one instance the outcome of the study (the dependent variable) cannot be measured at either the interval or ratio level, but the data can be ranked. Alternatively, the data are initially measured at the interval or ratio level but an assumption of the preferred statistical test, such as that there is a normal distribution, is violated and these data are then converted to ranks. Ranked data are encountered less commonly in the social sciences than either frequency or interval/ratio data. Accordingly, the **Kruskal-Wallis H test**, which is used to determine whether there is a difference when there are ordinal data, is reviewed in this appendix rather than in the main chapters of the text.

Kruskal-Wallis H test – An inferential procedure that is analogous to the one-way between-subjects ANOVA except that it is used with ordinal data.

One of the advantages of the procedures that employ ranked data is that they do not make as many assumption about the populations from which the samples are drawn as do the procedures utilized with interval or ratio data. More specifically, the Kruskal-Wallis H test does not estimate population parameters such as the mean or variance and, consequently, is called a nonparametric procedure. In addition, it does not assume that the samples are drawn from normally distributed populations, though the Kruskal-Wallis H test does assume that the populations have the same distributions. This greater flexibility is paired, however, with a loss of power and thus more subjects will be needed with this statistical test than would be needed with the procedures employed with interval or ratio data, such as the analysis of variance.

Though the equation that is used for the Kruskal-Wallis H test may at first seem peculiar, it is mathematically straightforward and is based upon a comparison of groups of ranked data. Fortunately, the basic logic of this test is similar to what was previously encountered with the one-way between-subjects ANOVA, which is used to analyze interval or ratio data. With the one-way between-subjects ANOVA, if the null hypothesis is true then the means of the scores in each of the treatment levels would be expected to be similar. This is because it would be unlikely for any treatment level, by chance, to have a preponderance of either high or low scores. Similarly, with the Kruskal-Wallis H test, which uses ordinal data, if the null hypothesis is true then we expect the means of the ranks in each of the treatment levels to be similar. Just as it is unlikely that a treatment level would consist mostly of high or low *scores* by chance, it is also unlikely that a treatment level would consist mostly of high or low *ranks* unless the treatment had the effect of changing the scores or ranks. With each procedure we then ascertain whether the observed outcome deviates enough from the expected outcome to reject the null hypothesis and accept the alternative hypothesis that the independent variable had an effect.

In Table A.1 you will see that the Kruskal-Wallis H test, which is underlined, is on the same *row* as the one-way between-subjects ANOVA. Thus, this test is appropriate when you have one independent variable with two or more independent samples. As was just noted, while the one-way between-subjects ANOVA is used when you have interval or ratio data, the Kruskal-Wallis H test is used with ordinal data or when results have been converted into ordinal data.

Table A.1 Overview of Inferential Statistical Procedures For Finding if there is a Difference

	Type of Data			
	Nominal (Frequency)		Ordinal (Ranked)	Interval/Ratio (Continuous Measure)
Research Design		Research Design		
One Variable With At Least Two Outcomes	Goodness-of-fit Chi-Square	One IV With One Sample		One-sample z Test or One-sample t Test
		One IV With Two Or More Independent Samples	Kruskal–Wallis H	One-way Between–Subjects ANOVA (Only two independent samples, Independent Samples t Test)
		One IV With One Sample Having Two Or More Repeated Measures		One-way Within–Subjects ANOVA (Only two repeated measures, Dependent Samples t Test)

Conducting The Kruskal-Wallis H Test With Ordinal Data

Assume we are interested in the order in which football players are chosen in NFL drafts. More specifically, we take a sample of 19 players and compare the order in which offensive players (Sample 1) were chosen with the order in which defensive players (Sample 2) were chosen and the order in which special teams players were chosen (Sample 3). The fictitious ranks, as well as totals needed for the computation of the Kruskal-Wallis H test, are shown in Table A.2. The null and alternative hypotheses are:

H_0 – The mean ranks for the three groups are the same.

H_1 – The mean ranks for the three groups are not the same.

We set α equal to .05.

Table A.2 Example 1: Order Players were Chosen

	Sample 1	Sample 2	Sample 3	
	1	2	12	
	3	4	14	
	6	5	16	
	8	7	17	
	9	11	18	
	10	15	19	
	13			
Total (T) =	50	44	96	
n =	7	6	6	N = 19

Note that the number of individuals in the samples do not have to be equal.

To calculate the Kruskal-Wallis H statistic, we use the following equation:

$$H = [\frac{12}{N(N+1)}][\Sigma(\frac{T^2}{n})] - 3(N+1)$$

where N = the total number of subjects, T = the total of the ranks for a sample, and n = the sample size.

This equation may be intimidating at first glance, but once you examine it carefully you will find it is actually quite easy to use.

The constants in the equation, 12 and 3, are not specific to this problem, they are part of the general equation for the Kruskal-Wallis H statistic. For our example, we would substitute and obtain:

$$
\begin{aligned}
\text{H} &= [\tfrac{12}{19\,(19\,+\,1)}][(\tfrac{50^2}{7}) + (\tfrac{44^2}{6}) + (\tfrac{96^2}{6})] - 3(19 + 1)\\
&= [\tfrac{12}{380}]\,[\,(\tfrac{2500}{7}) + (\tfrac{1936}{6}) + (\tfrac{9216}{6})\,] - 3(20)\\
&= [\,0.032\,]\,[\,357.143 + 322.667 + 1536.000\,] - 60\\
&= (0.032)(2215.810) - 60\\
&= 70.906 - 60\\
&= 10.91
\end{aligned}
$$

Note that these calculations were carried out to three decimal places before rounding to two places. This reduces the effects of rounding error.

The degrees of freedom for the Kruskal-Wallis H statistic are equal to the number of samples minus one. In our case, this would be 3 – 1, which equals 2. Referring to the chi-square table (Appendix K, Table 2), which is also used with the Kruskal-Wallis H test, we find a critical value of 5.99 with two degrees of freedom and alpha set at .05. As our obtained value of 10.91 is greater than the critical value we reject the null and accept the alternative hypothesis. In other words, we conclude that the mean ranks for the three groups are not the same.

We are still faced with two issues. First, we have not yet calculated a measure for effect size. Second, while the significant Kruskal-Wallis H statistic indicates that at least one of the sample's mean ranks is expected to differ from another sample's mean ranks, the test does not indicate which, or how many, of these samples' mean ranks differ. In other words, just as with the chi-square test of independence and the one-way between-subjects ANOVA, the Kruskal-Wallis H test provides an overall test of significance for the entire study, but when there are more than two samples it does not indicate where the significant difference(s) is (are). We will examine the issue of effect size first, and then describe a procedure for specifying where a difference within a significant Kruskal-Wallis H test is located.

Calculating The Effect Size

Eta squared (η^2) is a measure of effect size for the Kruskal-Wallis H test. As the following equation indicates, the effect size is easily found once the Kruskal-Wallis H statistic has been computed:

$$\text{Eta squared } (\eta^2) \text{ for the Kruskal-Wallis H test} = \frac{\text{H}}{\text{N} - 1}$$

where H is the value of the Kruskal-Wallis statistic and N is the total number of ranks.

$$= \frac{10.91}{19 - 1}$$

$$= 0.61$$

Eta squared (η^2) is an example of what is called a coefficient of determination, which was discussed in Chapter 14. A coefficient of determination indicates what proportion of variability in one variable is accounted for by the variability in another variable. In our case, η^2 equals 0.61. Therefore, 61% of the variability in the hypothetical rankings is accounted for by knowing whether the choice in the draft involves an offensive, defensive or special teams player.

Conducting The Post Hoc Comparisons

Next we return to the question of which sample ranks differ significantly. Based upon the result of the Kruskal-Wallis H test we rejected the null hypothesis and, therefore, expect that there exists at least one difference in the rankings between samples. However, we do not know where this (these) difference(s) may be. As with the chi-square test of independence and the one-way between-subjects ANOVA, we now need to perform post hoc tests in order to ascertain which specific comparisons are statistically significant. And, just as with the chi-square test and the one-way between-subjects ANOVA, we will simplify the situation by limiting ourselves to comparisons of the original samples and omit comparisons where samples are combined. We are, therefore, making what are called pairwise comparisons. There are k(k – 1) / 2 possible pairwise comparisons where k = the number of samples. In our example there are 3 independent samples and there would be 3(3 – 1) / 2, or 3 pairwise comparisons. These pairwise comparisons would be between Sample 1 and Sample 2, between Sample 1 and Sample 3, and between Sample 2 and Sample 3. Any one, any two, or all three of these comparisons could be statistically significant. Just as with the chi-square test and the one-way between-subjects ANOVA, a significant Kruskal-Wallis H test simply indicates that at least one comparison between pairs of samples is expected to be significant.

A number of post hoc procedures have been developed for use with the Kruskal-Wallis H test. The easiest alternative is to conduct a series of tests appropriate for use with a two-sample difference design and then utilize the Bonferroni method that was introduced in Chapter 8 to control the experimentwise error. As we have ordinal data, inspection of the overview table (Appendix L) indicates that our post hoc would be further Kruskal-Wallis H tests (there are other alternatives).

As you recall, the Bonferroni method, which was introduced in our discussion of the chi square test of independence in Chapter 8, maintains the experimentwise error by dividing the alpha level by the number of comparisons being made. In the present case there are three comparisons so we would divide our alpha of .05 by three to obtain .0167. Because this specific alpha level is not included in our chi-square table we turn to a table which includes more levels of alpha, or to a computer program. The definition of degrees of freedom remains the number of samples minus one. As we are now comparing the ranks of pairs of samples, the df are equal to 2 – 1. With an alpha of .0167 and 1 df our critical value becomes 5.73. We now proceed with our first post hoc comparison, a comparison of Sample 1 and Sample 2. For the post hoc we treat Sample 1 and Sample 2 as a complete set of data. The first step, therefore, is to re-rank the data for these two samples, as is shown in Table A.3:

Table A.3 Example 1: Post Hoc Analysis for Sample 1 and Sample 2

	Sample 1	Sample 2	
	1	2	
	3	4	
	6	5	
	8	7	
	9	11	
	10	13	
	12		
T =	49	42	
n =	7	6	N = 13

To calculate the Kruskal-Wallis H test we use the same equation as previously:

$$H = [\frac{12}{N(N+1)}][\Sigma(\frac{T^2}{n})] - 3(N+1)$$

For our example, we would substitute and obtain:

$$H = [\frac{12}{13\ (13+1)}][(\frac{49^2}{7}) + (\frac{42^2}{6})] - 3(13+1)$$

$$= [\frac{12}{182}]\ [(\frac{2401}{7}) + (\frac{1764}{6})] - 3(14)$$

$$= [\ 0.066\ ]\ [\ 343 + 294\ ] - 42$$

$$= (0.066)(637) - 42$$

$$= 42.042 - 42$$

$$= 0.04$$

As before, the initial calculations were carried out to three decimal places in order to reduce the effects of rounding error.

The re-ranking for the comparison between Sample 1 and Sample 3 is shown in Table A.4.

Table A.4 Example 1: Post Hoc Analysis for Sample 1 and Sample 3

Sample 1	Sample 3
1	7

	2	9	
	3	10	
	4	11	
	5	12	
	6	13	
	8		
T =	29	62	
n =	7	6	N = 13

To calculate the Kruskal-Wallis H test we would substitute these values and obtain:

$$H = [\frac{12}{13\ (13 + 1)}][(\frac{29^2}{7}) + (\frac{62^2}{6})] - 3(13 + 1)$$

$$= [\frac{12}{182}]\ [(\frac{841}{7}) + (\frac{3844}{6})] - 3(14)$$

$$= [\ 0.066\ ]\ [\ 120.143 + 640.667\ ] - 42$$

$$= (0.066)(760.810) - 42$$

$$= 50.213 - 42$$

$$= 8.21$$

The re-ranking for the comparison between Sample 2 and Sample 3 is shown in Table A.5.

Table A.5 Example 1: Post Hoc Analysis for Sample 2 and Sample 3

	Sample 2	Sample 3	
	1	6	
	2	7	
	3	9	
	4	10	
	5	11	
	8	12	
T =	23	55	
n =	6	6	N = 12

To calculate the Kruskal-Wallis H test we would substitute these values and obtain:

$$H = [\frac{12}{12\ (12 + 1)}][(\frac{23^2}{6}) + (\frac{55^2}{6})] - 3(12 + 1)$$

$$= [\frac{12}{156}]\ [(\frac{529}{6}) + (\frac{3025}{6})] - 3(13)$$

$$= [\ 0.077\ ]\ [\ 88.167 + 504.167\ ] - 39$$

$$= (0.077)(592.334) - 39$$

$$= 45.610 - 39$$
$$= 6.61$$

As the critical value, based upon the Bonferroni method, is 5.73 the comparison between the mean rank of Sample 1 and the mean rank of Sample 3, as well as the comparison between the mean rank of Sample 2 and the mean rank of Sample 3, are statistically significant. Inspection of these hypothetical data indicates that the special teams players were chosen later than either the offensive or defensive players were chosen. The comparison between the mean rank of Sample 1 and the mean rank of Sample 2 is not statistically significant. This indicates that the order in which offensive and defensive players were chosen did not significantly differ with these hypothetical data.

It is important to note that the Bonferroni method is quite conservative. This means that the probability of making a Type I error is somewhat less than the value the experimenter has chosen, in this case .05. A consequence is an increase in the probability of making a Type II error. In other words, while the Bonferroni method is very effective at preventing us from rejecting the null hypothesis when it is in fact correct (Type I error), it also increases the likelihood that we will fail to reject the null hypothesis when it is in fact false (Type II error). With a small number of comparisons the Bonferroni method is appropriate, but as the number of comparisons increases it becomes increasingly conservative and, therefore, there is an increased risk of making a Type II error. The rule of thumb is to use the Bonferroni method when there are five or fewer comparisons. If you need to conduct post hoc tests with a larger number of comparisons you should consult a more advanced statistics text to determine the appropriate procedure.

Reporting The Results Of A Kruskal-Wallis H Test

In a paper we would indicate that the overall statistical test was significant, provide our measure of effect size, and identify which specific group comparisons were found to differ. Specifically, we would provide the degrees of freedom, the total number of subjects, the calculated value, indicate that the overall Kruskal-Wallis H test was significant, and give the effect size ($H(2, N = 19) = 10.91$, $p < .05$, $\eta^2 = .61$). Further pairwise comparisons using the Bonferroni method indicated that the difference between the mean rank of Sample 1 and the mean rank of Sample 3 ($H(1, N = 13) = 8.21$, $p < .05$), as well as the difference between the mean rank of Sample 2 and the mean rank of Sample 3 ($H(1, N = 12) = 6.61$, $p < .05$), were statistically significant. The comparison between the mean rank of Sample 1 and the mean rank of Sample 2 was not statistically significant ($H(1, N = 13) = .04$, $p > .05$).

Use Of The Kruskal-Wallis H Test With Interval/Ratio Data

In our first example of the Kruskal-Wallis H test we utilized data that were collected as ranks. The Kruskal-Wallis H test is also commonly used with data that were originally collected at the interval or ratio level of measurement but which were then converted to ranks. Since interval or ratio data include more information than ranked data, the statistical tests that use the interval or ratio levels of measurement are more efficient than tests that rely upon levels of measurement that include less information. What this means is that a test that utilizes data at the interval or ratio level of measurement will, all else being equal, not need as many subjects in order to find a difference to be significant. Similarly, with ordinal data we do not need to collect data from as many subjects as we will have to if we use nominal data. So, you may ask, why would anyone convert interval or ratio data to ordinal data since this leads to a loss of information and, therefore, a loss of statistical power?

As was noted previously, the tests that utilize interval or ratio data are called parametric tests because they make assumptions about population parameters and they assume normal distributions. However, these assumptions may not be met. For instance, let us assume that we have collected the scores of a class exam. For this to be suitable for parametric analysis the data should be normally distributed. This means that there are many scores in the middle of the distribution and progressively fewer scores the farther we move, in either direction, from the middle. But what if most of the students in the class found the exam to be very easy? In this case, most students will have done well. We may get what is called a **ceiling effect**, with many scores clustered in the high 90s while fewer students scored lower. If so, the distribution will not be normal. Instead it will be negatively skewed. Particularly if the sample size is small it would not be appropriate to use a test that assumes normality even though the data are measured at the interval or ratio level. Instead, one option is to convert the scores to ranks and utilize a non-parametric procedure, such as the Kruskal-Wallis H test. It is important to remember, however, that while the Kruskal-Wallis H test does not assume the data come from normally distributed populations, it does assume the population distributions are the same. An example follows.

<u>Ceiling effect</u> – When the scores are predominately at the high end of the range of possible outcomes.

Let us assume that a physician is interested in how quickly two anesthesias take effect, measured in seconds. Each subject undergoing surgery is randomly assigned to receive one of the anesthesias. Since these are ratio data, the null and alternative hypotheses are:

H_0 – The mean of each sample is the same.

H_1 – The mean of each sample is not the same.

We set α equal to .05.

The time to reach a standard state of relaxation, in seconds, for each patient is indicated in Table A.6.

Table A.6 Example 2: Time for Two Anesthesias to Take Effect

Anesthesia A	Anesthesia B
12	15
13	17
14	17
20	26
24	36
29	42
36	72
61	
92	

Referring to Table A.1 or the overview table (Appendix L) will indicate that with two independent samples and ratio data we would initially consider using either the independent samples t test or the one-way between-subjects ANOVA. However, these are parametric tests that assume that the samples are drawn from populations that are normally distributed. This assumption is not met with our data as both samples are clearly positively skewed, and the sample sizes are small. It is, accordingly, inappropriate to use either an independent samples t test or an ANOVA with these data. Instead, we can convert the scores to ranks and conduct a nonparametric test such as the Kruskal-Wallis H test with two samples (there are other appropriate tests). The null and alternative hypotheses now become:

H_0 – The mean *rank* of each sample is the same.

H_1 – The mean *rank* of each sample is not the same.

The ranked data are presented in Table A.7, as well as totals needed for the computation of the Kruskal-Wallis H test. Note that when two scores are tied the mean of the two ranks involved is assigned to each of the scores. In our example, the scores for ranks 5 and 6 were tied in the second group so each was given the rank of 5.5. As the 5th and 6th scores have already been ranked, the next rank in the table is 7. In addition, the scores for ranks 11 and 12 are also tied so each is given the rank of 11.5. The next rank would, therefore, be 13. (If more than two scores were tied, essentially the same procedure would be followed. Each score would be given the same mean rank and the next assigned rank would reflect the number of ranks that had already been assigned.)

Table A.7 Example 2: Conversion of Data to Ranks

	Anesthesia A	Anesthesia B	
	1	4	
	2	5.5	
	3	5.5	
	7	9	
	8	11.5	
	10	13	
	11.5	15	
	14		
	16		
T =	72.5	63.5	
n =	9	7	N = 16

Now that the data are ranked, the Kruskal-Wallis H test is conducted as before:

$$
\begin{aligned}
\mathrm{H} &= [\frac{12}{\mathrm{N(N+1)}}][\Sigma\,(\frac{\mathrm{T}^2}{\mathrm{n}})\,] - 3(\mathrm{N}+1)\\
&= [\frac{12}{16\,(16+1)}][(\frac{72.5^2}{9}) + (\frac{63.5^2}{7})] - 3(16+1)\\
&= [\frac{12}{272}]\,[\,(\frac{5256.25}{9}) + (\frac{4032.25}{7})] - 3(17)\\
&= [\,0.044\,]\,[\,584.028 + 576.036\,] - 51\\
&= (0.044)(1160.064) - 51\\
&= 51.043 - 51\\
&= 0.04
\end{aligned}
$$

The degrees of freedom are equal to the number of samples minus one. In our case this would be 2 – 1 = 1. Referring to the chi-square table (Appendix K, Table 2), we find a critical value of 3.84 with alpha set at .05. As our obtained value of 0.04 is less than the critical value we retain the null hypothesis that there is no difference in the mean rank of the two anesthesias. (If this Kruskal-Wallis H test was statistically significant we would determine the effect size, but we would not conduct a post-hoc test since there are only two groups being compared and thus we already know which groups differ.)

Reporting The Results Of A Kruskal-Wallis H Test

In a paper we would write that the Kruskal-Wallis H test was not significant (H(1, N = 16) = 0.04, p > .05).

Purpose And Limitations Of Using The Kruskal-Wallis H Test

1. *This is a test for equality of ranks.* The null hypothesis is that the mean rank for each sample is equal. In other words any difference in the distribution of the ranks is due to chance.
2. *This is an overall test of significance.* In designs with more than two samples a statistically significant outcome indicates that a difference in the rankings exists between the samples, but the initial Kruskal-Wallis H test does not indicate where that difference(s) is (are). Post hoc tests would need to be conducted to identify the specific groups that differ.
3. *The test does not provide a measure of effect size.* The Kruskal-Wallis H test is a test of significance. It indicates whether or not an outcome is likely to have occurred by chance. If the Kruskal-Wallis H statistic is significant, a measure of effect size, such as eta squared, should then be calculated.
4. *The sample size should not be too small.* As a general rule, no sample should have fewer than 5 subjects.

Assumptions Of The Kruskal-Wallis H Test

1. *You have ordinal data.* The data are in the form of ranks or can be converted into ranks.
2. *The observations are independent.* Each subject provides only one datum and no subject is matched with another subject during assignment to samples.
3. *Population distributions are the same.* When used with data originally collected at the interval or ratio levels the Kruskal-Wallis H test does not require that populations are normally distributed. However, it does require that the population distributions from which the samples were drawn are the same.

Conclusion

We have now completed our introduction to the Kruskal-Wallis H statistic. Before moving on to other tests it might be valuable to review how the Kruskal-Wallis H statistic is related to other procedures. By referring to Table A.1, you will see that the Kruskal-Wallis H test is on the same *row* as the one-way between-subjects ANOVA. Each of these tests is used with a different level of measurement, the Kruskal-Wallis H test for ordinal data and the ANOVA for interval or ratio data. In other respects they are quite similar. Each is used when you have a design with one IV and two or more independent samples, each provides an overall test of whether there is a difference but with more than two samples they do not indicate what the specific basis of this difference is and, if significant, each should be followed by a measure of effect size.

The Kruskal-Wallis H test is also on the same row of Table A.1 as the independent samples t test. Like the ANOVA, the independent samples t test utilizes interval or ratio data. However, the t test is limited to designs in which there are only two levels of the independent variable. The Mann-Whitney U test is usually given as the direct parallel to the independent samples t test when there are ordinal data since it is also limited to situations with two levels of the independent variable. However, just as the more flexible one-way between-subjects ANOVA can be utilized instead of the independent samples t test, the more flexible Kruskal-Wallis H test can be utilized instead of the Mann-Whitney U test.

Glossary Of Terms

Ceiling effect – When the scores are predominately at the high end of the range of possible outcomes.

Kruskal-Wallis H test – An inferential procedure that is analogous to the one-way between-subjects ANOVA except that it is used with ordinal data.

Questions – Appendix A

(Answers are provided in Appendix J.)

1. The Kruskal-Wallis H test is used with _____ data.
    a. Nominal
    b. Ordinal
    c. Interval/ratio
    d. It can be used with any scale of measurement.

2. The Kruskal-Wallis H test is appropriate when you have _____.
    a. two or more independent samples
    b. repeated measures
    c. one group compared to a known population parameter
    d. one independent sample and one repeated measures sample

3. The degrees of freedom for the Kruskal-Wallis H statistic are equal to _____.
    a. The number of subjects minus one
    b. The number of subjects minus the number of samples
    c. The number of samples plus the number of subjects
    d. The number of samples minus one

4. The table used with the Kruskal-Wallis H test is also used with the _____.
    a. ANOVA
    b. Independent samples t test
    c. Chi-square test
    d. z score

5. The measure of effect size for the Kruskal-Wallis H test is _____.
   a. Eta squared (η^2)
   b. The square of r
   c. Not defined
   d. Difficult to calculate

6. The _____ indicates what proportion of variability in one variable that is accounted for by the variability in another variable.
   a. Value of the Kruskal-Wallis H test
   b. Number of the degrees of freedom minus one
   c. Coefficient of determination
   d. Coefficient of nondetermination

7. Following a significant overall Kruskal-Wallis H test with three or more samples, further pairwise comparisons using the _____ for control would be used.
   a. Bonferroni method
   b. Tukeys HSD
   c. Chi-square
   d. Any of the above

For questions 8 – 11 use the following information: A banker wanted to compare the incomes of people living in two sections of a city. To do so he conducted a questionnaire study. He noted that the ratio data were positively skewed and thus converted them to ranks. The ranks are:

Group 1	Group 2
2	1
3	6
4	9
5	10
7	13
8	14
11	15
12	16
	17

8. What is the value of the H test?
   a. 1.774
   b. 2.673
   c. 3.68
   d. 4.51

9. What is the critical value (alpha = .05, two tailed test)?
   a. 1. 96
   b. 2.58
   c. 3.16
   d. 3.84

10. Do the ranks of the two groups differ statistically?
    a. Yes
    b. No

11. What is the value of eta squared?
 a. .212
 b. .28
 c. Eta squared should not be calculated as the H test was not significant

Problems 12 – 15 examine the effect of making one switch (which is underlined) in the rankings of each of the above groups:

Group 1	Group 2
2	11
3	6
4	9
5	10
7	13
8	14
1	15
12	16
	17

12. What is the value of the H test?
 a. 1.77
 b. 2.67
 c. 3.76
 d. 7.99

13. What is the critical value (alpha = .05, two tailed test)?
 a. 1.96
 b. 2.58
 c. 3.16
 d. 3.84

14. Do the ranks of the two groups differ statistically?
 a. Yes
 b. No

15. What is the value of eta squared?
 a. .21
 b. .50
 c. Eta squared should not be calculated as the H test was not significant

Appendix B
Phi Correlation:
Identifying the Strength of an Association when there are Nominal Data

"You can't fix by analysis what you bungled by design."

Light, Singer and Willett, page viii

Introduction

The **phi correlation** is a commonly used statistic when there are nominal data. The Greek letter phi is ϕ. Thus the symbol ρ_ϕ would indicate the population correlation, and r_ϕ would signify that we are dealing with samples, though these symbols, as well as phi r, are rarely employed. Instead the phi correlation is usually represented as simply ϕ. However, in order to make it clear that the phi correlation is a type of Pearson r, I will use r_ϕ as the symbol in this appendix. The phi correlation is located on the same row of Table B.1 as the Pearson r, and is underlined in the table.

Phi correlation (r_ϕ) – A form of Pearson correlation used with nominal data when both variables are dichotomous.

Table B.1 Overview of Statistical Procedures for Association Studies

	Type of Data		
	Nominal (Frequency)	Ordinal (Ranked)	Interval/Ratio (Continuous Measure)
Research Question			
Association:	Chi-Square Test of Independence		
Correlation:	Phi r	*Spearman r*[a]	Pearson r *Multiple Correlation*[b]
Regression:			Regression *Multiple Regression*[b]

Italicized items are reviewed in the following appendixes:

a. Appendix C
b. Appendix D

It is important to note that with phi r both variables must be dichotomous. In other words, each variable is limited to two, mutually exclusive options. For example, you either passed a particular course, or you didn't. And you have either visited Ireland, or you haven't.

Conducting The Phi Correlation

You are familiar with phi as a measure of effect size from the discussion in Chapter 8, but we will now be utilizing phi as a correlation. In Chapter 8 we calculated a 2 X 2 chi-square statistic to test whether there was a *difference* or an *association* between men's and women's views of infidelity. (Recall that with nominal data the distinction between finding a difference versus finding an association is less clear than is the case with ordinal, interval or ratio data.) Table 8.9, which includes the marginal totals, is reproduced below in a slightly modified form as Table B.2. The four cells in the Table are labeled a, b, c, and d. As the chi-square was found to be statistically significant, we rejected the null hypothesis that there was no difference (association) in the distribution of answers for men and women and accepted the alternative hypothesis that these distributions do differ (are associated). More specifically, we concluded that men were more distressed by sexual infidelity whereas women were more distressed by emotional infidelity. This statistically significant difference (association) indicates that the outcome was unlikely to have occurred by chance. We then used phi as a measure of effect size.

Table B.2 An Example of the Phi Correlation

	Women	Men	Total
More distressed by emotional infidelity	42 (a)	12 (b)	54 (a + b)
More distressed by sexual infidelity	17 (c)	48 (d)	65 (c + d)
Total	59 (a + c)	60 (b + d)	

If, instead, we viewed this as a correlational study our focus would now shift to determining the strength of the relationship. And to do so, we would not begin by calculating a chi-square. Instead, we would calculate the phi correlation directly using the following equation:

$$r_\phi = (ad - bc) / \sqrt{[(a + b)(c + d)(a + c)(b + d)]}$$

In our case:

$$r_\phi = (42 \text{ X } 48) - (12 \text{ X } 17) / \sqrt{[(54)(65)(59)(60)]}$$

$$= (2016 - 204) / \sqrt{[(54)(65)(59)(60)]}$$

$$= 1812 / 3524.97$$

$$= .51$$

Note that this is the same value that we calculated for phi when used as a measure of effect size in Chapter 8. However, now we calculated r_ϕ directly without first calculating a chi-square value.

(If the value of r_ϕ is negative, change it to positive. The sign simply indicates the order of the columns and rows. For instance, if the proportions for men had been listed in Table B.2 before the proportions for women, r_ϕ would have been negative. Thus the sign is not meaningful and can be ignored.)

As we are now dealing with a correlation, the null hypothesis, H_0, states that there is no relationship between the two variables. In other words, if H_0 is true, the obtained value of r_ϕ should not differ significantly from 0. The alternative hypothesis, H_1, states that there is a relationship between the two variables. In other words, if H_1 is true, the obtained value of r_ϕ should differ significantly from 0. To test whether the obtained value of r_ϕ, in this case .51, is significantly different from 0, we convert r_ϕ into a chi-square and then turn to the chi-square table. To do so, we use the following equation:

Chi-square = $(n)(r_\phi)^2$ where n = the total number of observations

= a + b + c + d

In our case:

Chi-square = $(42 + 12 + 17 + 48)(.51)^2$

$= (119)(.51)^2$

$= (119)(.26)$

$= 30.94$

Note that this value for the chi-square is, except for rounding error, the same value that was obtained when the chi-square was calculated directly for these data in Chapter 8.

The df are equal to:

(number of columns – 1) X (number of rows – 1). Since phi r deals with two dichotomous variables, we have a 2 X 2 table. The df are thus equal to (2 – 1)(2 – 1) = 1.

If the obtained value for this chi-square is greater than the critical value listed in the chi-square table, we reject the null hypothesis and accept the alternative. With 1 df the critical value is 3.84. Our calculated chi square is, therefore, significant at the .05 level, and is beyond the critical value for the chi-square with 1 df even at alpha equal to .01. (It is actually beyond the critical value with alpha equal to .005.) We therefore reject the null hypothesis and accept the alternative that the obtained value of phi r is significantly different from 0.

As you recall, a correlation such as the Pearson r can vary from –1 to +1. However, r_ϕ only meaningfully varies from 0 to 1, where 0 indicates that there is no relationship between the two variables and 1 indicates there is a perfect correlation between the two variables. The larger r_ϕ is, the better we can predict.

The strength of the association can best be illustrated by finding the square of the correlation. Recall that the square of a correlation is called the coefficient of determination. In our case it would be $(r_\phi)^2$. The coefficient of determination measures what proportion of variance in one variable is explained or accounted for by the other variable. In our example, the correlation was equal to .51. The coefficient of determination ($r_\phi{}^2$) is, therefore, equal to $.51^2$, which is .26 or 26%. This indicates that knowing whether a subject is a man or a woman will remove or account for 26% of the variability in predicting their view of whether sexual or emotional infidelity is more distressing.

And, just as with the Pearson r, we can also calculate a coefficient of nondetermination, which with r_ϕ would be equal to $1 - r_\phi{}^2$. In our example this would be $1 - .51^2$, which equals 1 – .26, which is .74. Thus, 74% of the variability in the response to infidelity is not accounted for by knowing whether the subject is a man or woman.

Reporting The Results Of A Phi Correlation

In order to provide a complete report of our finding, we would say that there was a significant correlation between the gender of the subject and their view of infidelity ($\phi = .51$, $p < .005$). The coefficient of determination, $r_\phi{}^2$, equaled .26. With this statement, we have indicated to the reader that a phi correlation was conducted, the size of the correlation and that it was statistically significant. Finally, we have provided a measure of the strength of the association to assist the reader in interpreting the size of the effect.

Remember, phi r is a correlation. It provides a measure of the magnitude of the relationship. It does not indicate that this is a causal relationship.

Purpose And Limitations Of Using The Phi Correlation

1. **Provides a measure of the strength of the association of two dichotomous variables.** The phi correlation provides a measure of the degree to which two dichotomous variables are related.
2. **Not a measure of cause and effect.** Phi r (r_ϕ) is a type of correlation. Correlational designs lack the level of experimenter control that is needed in order to justify coming to a cause and effect conclusion. Thus, the researcher cannot conclude that one variable caused a change in the other.

Assumptions Of The Phi Correlation

1. **Nominal data.** The data are in the form of frequencies or can be converted to frequencies.

2. **Data are in the form of two dichotomies.** The phi correlation is used when there are two dichotomous variables.

Conclusion

The phi correlation is an easy to calculate measure of the strength of the relationship between two dichotomous variables, each consisting of nominal data. It is a form of Pearson r and, as with the Pearson r, the square of the phi correlation provides a useful measure of effect size. However, remember that unlike the Pearson r, the sign of the phi correlation is always reported as being positive.

Glossary Of Terms

<u>Phi correlation (r_ϕ)</u> – A form of Pearson correlation used with nominal data when both variables are dichotomous.

Questions – Appendix B

(Answers are provided in Appendix J.)

1. If you are interested in whether two variables are correlated and if both variables are nominal and are dichotomous (a score falls in one group or another) then you would use _____.
   a. Chi-square
   b. Phi r
   c. Spearman r
   d. Pearson r

2. If you are interested in whether two variables are correlated and if both variables consist of ordinal data you use _____.
   a. Chi-square
   b. Phi r
   c.. Spearman r
   d. Pearson r

3. If you are interested in whether two variables are correlated and if you are dealing with two interval or ratio variables you would employ _____.
   a. Chi-square
   b. Phi r
   c. Spearman r
   d. Pearson r

4. In the case of the _____ correlation, the sign simply reflects the order that the variables were entered into an equation. If the order is reversed, the sign will also change.
   a. Phi r
   b. Spearman r

c. Pearson r

5. With a (an) _____ design, we are not asking if the distributions or proportions that we have observed are different. Instead, we are asking if the variables are related.
a. Experimental
b. Descriptive statistical
c. Association
d. ANOVA

6. The _____ indicates the proportion of variance in one variable that is explained or accounted for by the other variable.
a. Spearman correction
b. coefficient of determination
c. co-variance
d. regression equation

7. Subjects are asked whether they enjoy watching baseball, and then are asked if they enjoy watching football. The following hypothetical data are obtained. What is the correlation?

		Like Baseball	
		Yes	No
Like Football	Yes	15	9
	No	5	12

a. .23
b. .33
c. .37
d. .52

8. Subjects are asked whether they like chicken wings, and then whether they like pizza. The following hypothetical data are obtained. What is the correlation?

		Like Pizza	
		Yes	No
Like Chicken Wings	Yes	12	4
	No	5	8

a. .23
b. .33
c. .37
d. .52

Appendix C
Spearman Correlation:
Identifying the Strength of an Association when there are Ordinal Data

"The fewer the facts, the stronger the opinion."

Arnold H. Glasow

Introduction

The **Spearman correlation**, also sometimes called the Spearman r or Spearman rank order correlation coefficient, is a commonly used statistic. It is located on the same *row* of Table C.1 as Phi r and the Pearson r, and is underlined in the table. Each of these correlations provides a measure of the strength of the association between two sets of numbers. Thus, when using correlations we are not asking if samples differ. Instead, we are asking if the samples are related and, if so, the strength of this relationship.

Spearman correlation (r_S) – A form of Pearson correlation used when the two variables are measured at the ordinal level.

Table C.1 Overview of Statistical Procedures for Association Studies

	Type of Data		
	Nominal (Frequency)	Ordinal (Ranked)	Interval/Ratio (Continuous Measure)
Research Question			
Association:	Chi-Square Test of Independence		
Correlation:	Phi r[a]	Spearman r	Pearson r *Multiple Correlation*[b]
Regression:			Regression *Multiple Regression*[b]

Italicized items are reviewed in the following appendixes:

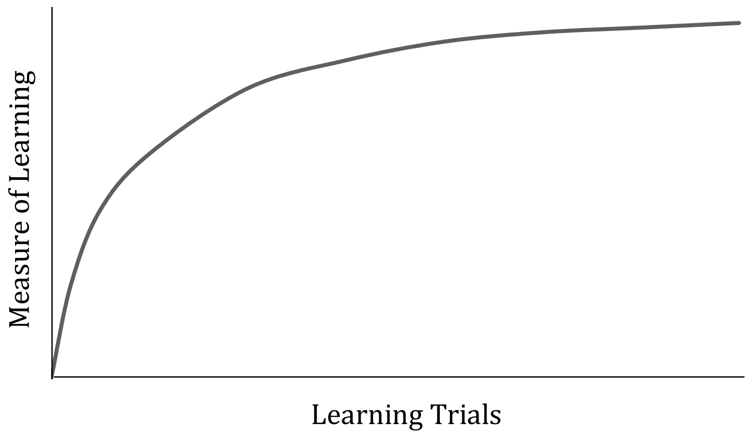
a. Appendix B

b. Appendix D

With the Spearman r we employ symbols that are similar to those that were used for the Pearson correlation. More precisely, the symbol ρ_S indicates the population correlation and r_S signifies that we are dealing with samples.

In the case of the Spearman r both variables will have been measured at the ordinal level or, more commonly, will have been converted from an interval or ratio level into ordinal data. Conversion into an ordinal level is undertaken because an assumption of the preferred correlation for interval or ratio data, the Pearson r, has not been met. For instance, the Pearson r is a measure of linear relationship. It only provides an accurate measure if the two interval or ratio variables have a straight-line relationship between them. However, many relationships, such as the classic example of a learning curve in which learning occurs rapidly at first, but then slows, are not linear (Figure C.1). If a Pearson r is calculated for these data it will underestimate the true degree of the relationship.

Figure C.1 Example 1: Example Of A Nonlinear Relationship – A Learning Curve



One solution would be to convert the ratio data into ordinal data and then conduct a Spearman correlation. This will provide a measure of the consistency of the rankings between your two variables. In other words, with the Pearson correlation we are asking “as one variable gets larger, does the other variable either increase or decrease in a straight-line fashion”? With the Spearman correlation we drop the requirement for a straight-line relationship. We are simply interested in knowing whether there is a reliable or consistent relationship between the order of the changes in the two variables.

An example of converting ratio scores into ordinal data and then using the Spearman r comes from the history of aviation. Before the advent of jet engines, airplanes were powered by piston engines, as are most cars today. Table C.2 provides the horsepower and the maximum speed (measured in miles per hour) for a number of historically significant airplanes.

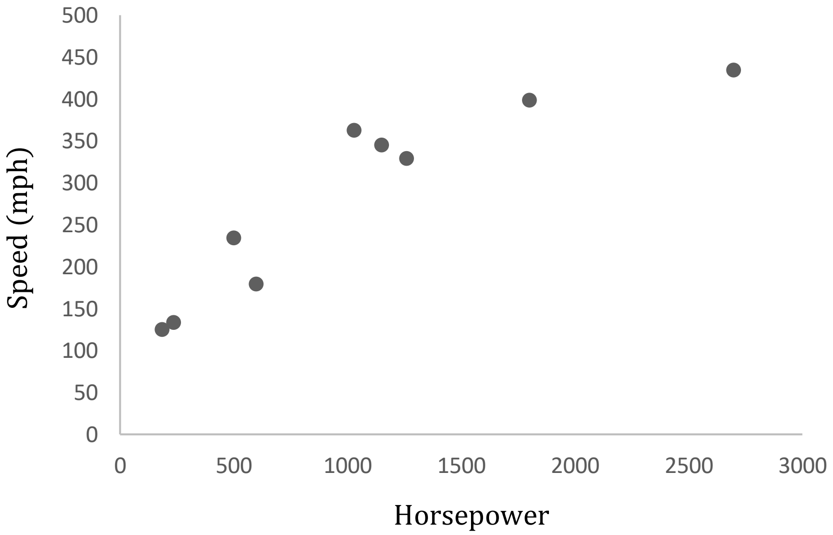
Table C.2 Example 2: Horsepower and Maximum Speed of Select Planes

Type of Plane	Horsepower	Speed (mph)
Fokker D VII	185	125
Spad XIII	235	133
Boeing P–26	500	234
Curtis P–6A Hawk	600	179
Spitfire IA	1030	362
P–40 C	1150	345
Hurricane II C	1260	329
FW–190 A–6	1800	398
F8F Bearcat	2700	434

Note. mph, miles per hour (1 mile = 1.60 kilometers).

Figure C.2 is a graph of the relationship between horsepower and speed.

Figure C.2 Example 2: A Nonlinear Relationship for Horsepower and Maximum Speed



It is evident that as the horsepower increases, so does the speed. However, it is also evident that with increased horsepower a plateau is reached. In other words, there is not a linear relationship between power and speed. It would, therefore, be inappropriate to calculate a Pearson r for these data. Instead, we can convert the data in Table C.2 into ranks, as shown in Table C.3, and then conduct a Spearman correlation.

The null and alternative hypotheses are:

H_0 – There is no association between the rank of horsepower and the rank of speed.

H_1 – There is an association between the rank of horsepower and the rank of speed.

As usual we set α equal to .05.

Table C.3 Example 2: Ranks of Horsepower and Maximum Speed of Select Planes, Along with Initial Calculations

Type of Plane	Horsepower	Speed (mph)	Difference (D)	D^2
Fokker D VII	1	1	0	0
Spad XIII	2	2	0	0
Boeing P–26	3	4	–1	1
Curtis P–6A Hawk	4	3	1	1
Spitfire IA	5	7	–2	4
P–40 C	6	6	0	0
Hurricane II C	7	5	2	4
FW–190 A–6	8	8	0	0
F8F Bearcat	9	9	0	0
			$\sum D = 0$	$\sum D^2 = 10$

Note. mph, miles per hour (1 mile = 1.60 kilometers).

As a check on our calculations, the sum of the differences between the pairs of ranks (the column headed by 'D') should always equal zero.

The equation for the Spearman r is:

$$r_s = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)}$$

where $\sum D^2$ is the sum of the squared differences between pairs of ranks and n is the number of *pairs* of data, in our case 9. The number 6 is a constant in the equation and is not related to our specific example.

Substituting, we find:

$$\text{Spearman r} = 1 - \frac{6\,(10)}{9\,(9^2 - 1)}$$

$$= 1 - \frac{60}{9\,(81 - 1)}$$

$$= 1 - \frac{60}{9\,(80)}$$

$$= 1 - \frac{60}{720}$$

$$= 1 - 0.08$$

$$= .92$$

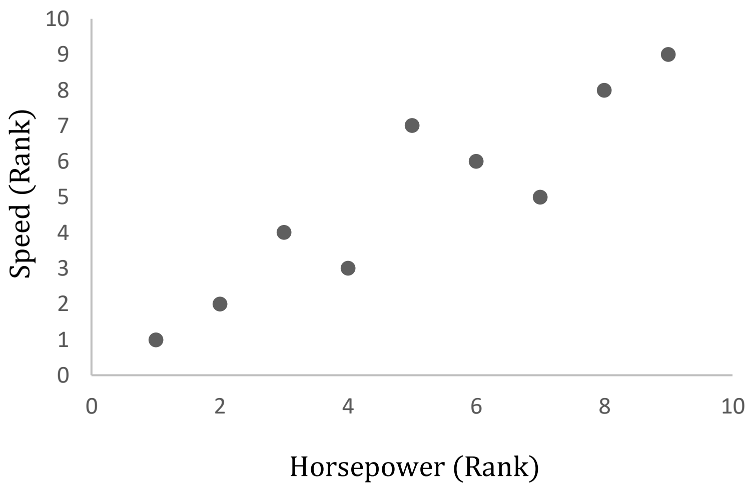
We now consult the Spearman r table (Appendix K, Table 7) to determine if the Spearman correlation that we just calculated is statistically different from 0. There are two characteristics of this table that differ from most other tables of critical values. First, *the Spearman r table is not based upon degrees of freedom but rather is based directly upon n, in this case the number of pairs of data*. Thus, we do not calculate degrees of freedom before consulting the Spearman r table. And second, *to be significant, the obtained value must be equal to or greater than the value listed in the table*. With most other statistical procedures the obtained value has to be greater than the critical value in the table.

For our example we refer to the row for an n of 9. With alpha equal to .05, we would need an outcome equal to or greater than .70 if we did not specify a predicted direction for the relationship before the data were collected, for this would be a two-tailed test. However, if we had specified that increased horsepower would lead to increased speed before the data were collected this would be a one-tailed test and the critical value would instead be .60. (Only the table for two-tailed tests is included in the appendixes.)

An appropriate time to have used the one-tailed test would have been with our airplane data. After all it would have been reasonable to assume before any data were collected that more powerful engines would have resulted in faster planes. However, we did not specify this direction before we collected the data so we will use the two-tailed critical value of .70. As our obtained value is .92 we reject the null that there is no relationship between power and speed and accept the alternative that there is a relationship. In fact, this outcome is significant at even the .01 level (critical value would be .83).

Inspection of the data indicates that as the engine power increases, so does the speed. This is also what the positive sign of the correlation indicates. When a Spearman correlation has a positive sign it indicates that as the ranks for one variable increase, so do the ranks for the other variable. A graph of our ranks for the horsepower and speed of planes is shown in Figure C.3. Note that if a Spearman correlation is positive it also indicates that as the ranks of one variable *decrease*, so do the ranks of the other. In other words, with a positive correlation the graph will rise to the right, and fall to the left.

Figure C.3 Example 2: Relationship of the Ranks of Aircraft Horsepower and Maximum Speed



It is also important to note that in the Spearman r table there are no negative numbers. The table indicates how large a Spearman r correlation must be in order to reject the null hypothesis *without regard to its sign*. Thus, if we had found a correlation of –.92, we would ignore the sign and compare .92 with the critical value found in the table. Of course, with a two-tailed test, the sign doesn't matter in determining whether to reject the null hypothesis. However, with a one-tailed prediction the sign does matter and the researcher would then have to check that the outcome was in the predicted direction.

To this point, we have taken a set of ratio data, converted it to ranks, calculated a Spearman r, and determined that it is statistically significant. Just as with the Pearson r we would also like to have a measure of the strength of the association in order to help us interpret what our significant outcome indicates. With the Spearman r, as with the Pearson r, a commonly used measure of effect size is the coefficient of determination. It is simply the square of the correlation, in this case, r_S^2.

With the Spearman correlation, the coefficient of determination indicates how much of the variability in one set of ranks is explained by the variability in the other set of ranks. In our case, the Spearman r equals .92 and r_S^2 equals .85. Thus, by knowing the rank of the aircraft's engine power we have explained 85% of the variability in the rank of the aircraft's speed. In the social sciences any coefficient of determination that is greater than .25 would be considered to be large.

Put another way, our analysis indicates that there is only 15% of the variability in the ranks of the aircrafts' speeds that is *not* accounted for by knowing the ranks of the aircrafts' horsepower (1.00 – 0.85 = 0.15 or 15%). As was discussed with the Pearson r, this value, which is the proportion of the variability of one variable not explained or accounted for by the variability of the other variable, is called the coefficient of nondetermination. For the Spearman r it is equal to $1 - r_S^2$.

Reporting The Results Of A Spearman Correlation

In order to report our finding, we would say that there was a significant correlation between the ranked horsepower and ranked speed of a series of aircraft ($r_S(9) = .92, p < .01$). The coefficient of determination, r_S^2, equaled .85. With this statement, we have indicated to the reader that the data were ranked, the number of pairs of scores, the size of the Spearman r correlation and that it was statistically significant. Finally, we have provided a measure of the strength of the association, or effect size, to assist the reader in interpreting the size of the effect.

Purpose And Limitations Of Using The Spearman Correlation

1. *Provides a measure of the association of two ranked variables.* The Spearman r provides a measure of the strength and direction of an association between two ranked variables.
2. *Not a measure of cause and effect.* The Spearman r is a type of correlation. Due to a lack of control in a correlational design a researcher is not justified in coming to a cause-and-effect conclusion.

Assumptions Of The Spearman Correlation

1. *Ordinal data.* The data are in the form of ranks or have been converted to ranks.
2. *Data are paired.* The data come as pairs, usually two measures on the same individual.
3. *No tied ranks.* The Spearman r correlation assumes that there are no tied ranks. If there are only a few the Spearman r will remain reasonably accurate. However, if there are a substantial number of tied ranks the Spearman r should not be employed.

Conclusion

The Spearman correlation is a commonly utilized, easy to calculate, measure of the relationship between two ranked variables. It is a form of Pearson r and like the Pearson r the sign of the Spearman r indicates the direction of the relationship and the square of the Spearman r provides a measure of effect size.

Glossary Of Terms

Spearman correlation (r_S) – A form of Pearson correlation used when the two variables are measured at the ordinal level.

Questions – Appendix C

(Answers are provided in Appendix J.)

1. The Spearman correlation is used with _____ data.
   a. Nominal
   b. Ordinal
   c. Interval/ratio
   d. Any of the above

2. The Spearman r _____ assume a linear relationship between the variables.
   a. Does
   b. Does not

3. Following the calculation of the Spearman r we use _____ when referring to the table to determine if the outcome is significant.
   a. The sample size, n
   b. df
   c. n – 1
   d. we do not refer to a table to determine the significance of the outcome

4. If a Spearman r is positive, _____.
   a. as the ranks of one variable increase, so do the ranks of the other
   b. as the ranks of one variable decrease, so do the ranks of the other
   c. as the ranks of one variable increase, the ranks of the other decrease
   d. both ‘a’ and ‘b’, but not ‘c’

5. With the Spearman r a commonly used measure of effect size is _____.
   a. The coefficient of determination
   b. The square root of the correlation
   c. r_S^2
   d. Both ‘a’ and ‘b’ but not ‘c’
   e. Both ‘a’ and ‘c’ but not ‘b’

6. The proportion of the variability of one variable not explained or accounted for by the variability of the other variable, is called the _____.
   a. coefficient of determination
   b. coefficient of nondetermination
   c. r_S^2
   d. both ‘b’ and ‘c’

7. If there are a substantial number of tied ranks, the Spearman r _____.
   a. should not be employed
   b. can still be employed
   c. will become equivalent to a Pearson r

Questions 8 – 11 deal with the relationship between studying and grades. Specifically, a professor gave an exam and asked how many hours students had studied for it. The data were:

Student	Hours Studied	Exam Grade
1	12	96
2	6	94
3	6	90
4	5	85
5	3	80
6	2	71
7	4	68

8. What is the Spearman r for these data?
    a. .44
    b. .74
    c. .88
    d. .97

9. What is the critical value assuming alpha equals .05 and this was a two-tailed test?
    a. .65
    b. .70
    c. .74
    d. .79
    e. .89

10. Is the outcome statistically significant?
    a. yes
    b. no

11. What is the coefficient of determination?
    a. .78
    b. .71
    c. .84
    d. .66

Appendix D
Multiple Correlation and Regression

"The difficulty lies, not in the new ideas, but in escaping the old ones …"

John Maynard Keynes

Introduction

Multiple correlation and multiple linear regression (usually called multiple regression) were introduced in Chapter 14. Multiple correlation (underlined in Table D.1) is a commonly utilized extension of the Pearson correlation, and multiple regression (underlined in Table D.1) is a commonly used extension of linear regression. In simple linear regression a statistically significant Pearson correlation is followed by the determination of the equation for the regression line. This equation, which can be written as $\hat{Y} = bX + a$, permits the prediction of Y (the criterion variable) from X (the predictor variable). It is defined as the straight line such that the sum of the squared errors of estimate is a minimum. So long as the Pearson correlation is statistically significant predictions based upon the regression equation will be more accurate than simply choosing the mean of Y for all values of X. In other words, the standard error of estimate will be less than the standard deviation. (Refer to Chapter 14 for a review.)

Table D.1 Overview of Statistical Procedures for Association Studies

	Type of Data		
	Nominal (Frequency)	Ordinal (Ranked)	Interval/Ratio (Continuous Measure)
Research Question			
Association:	Chi-Square Test of Independence		
Correlation:	*Phi r*[a]	*Spearman r*[b]	Pearson r Multiple Correlation
Regression:			Regression Multiple Regression

Italicized items are reviewed in the following appendixes:

a. Appendix B
b. Appendix C

The logic of correlation and simple linear regression (also called linear regression) can be extended to situations in which there is more than one predictor of Y. For example, if we were interested in predicting students' college grade point averages (Y) based upon their high school grade point averages (X), we would use correlation and linear regression assuming the appropriate assumptions were met. However, there are other reasonable predictors of college grade point average including a student's SAT score and a measure based upon their letters of recommendation. With *multiple correlation* we can determine the strength of the association for this set of predictors with college grade point average. And with *multiple linear regression* we can combine these predictors into one equation to estimate the student's college grade point average. By doing so our ability to predict Y is likely to be improved. The equation for multiple regression with two predictors can be written in a number of ways. One form of the equation is as follows:

$$\hat{Y} = B_1X_1 + B_2X_2 + B_0$$

where $\hat{Y}$ is the predicted value of the criterion variable; B_1 is the regression weight associated with the first predictor variable, X_1; B_2 is the regression weight associated with the second predictor variable, X_2; and B_0 is a constant, the value of the Y intercept (the value of Y when the regression line crosses the Y axis).

The equation for multiple regression with three predictors could be written as:

$$\hat{Y} = B_1X_1 + B_2X_2 + B_3X_3 + B_0$$

where each term is defined as before and where B_3 is the regression weight associated with the third predictor variable, X_3.

While these equations may appear 'new', they are actually just more involved forms of the equation we employed for linear regression, $\hat{Y} = bX + a$. As we are now dealing with multiple regression, and thus additional predictors, the equation has additional terms. Two other things to note are that while the slope of the line in linear regression (also called the regression weight) is indicated by the letter 'b' in the equation $\hat{Y} = bX + a$, in multiple regression an equivalent regression weight is often signified with a capital letter 'B' along with a subscript '1', '2', etc. And, while the Y intercept in linear regression is indicated by the letter 'a', in multiple regression it is indicated by the symbol B_0. Thus, while the multiple regression equation is more complex than the linear regression equation it should be evident that it is a straightforward extension.

There is no theoretical limit to the number of possible predictors in multiple regression. However, the computations quickly become tedious and thus anyone who will be utilizing multiple regression is strongly encouraged to employ a computer-based statistical package. Finally, while there is only one type of linear regression there are a number of approaches to multiple regression.

In **simple multiple regression** all of the predictor variables are assessed simultaneously without any consideration of theoretical importance or prior findings. Any predictors that do not significantly enhance the overall prediction are dropped.

With **hierarchical multiple regression** the researcher specifies the order in which predictor variables are entered into the regression equation. Only those variables that provide a statistically significant improvement in the ability to predict Y as they are entered are maintained in the equation.

In **forward stepwise multiple regression** predictor variables are entered into the regression equation based upon their individual correlations with Y. Thus, the predictor variable with the strongest individual correlation with Y is entered first. Then the variable that accounts for the greatest proportion of the remaining variability is entered. This process continues until there is no longer a statistically significant improvement in the ability to predict Y as variables are entered.

With **backward stepwise multiple regression** all of the predictor variables are assessed simultaneously as in simple multiple regression. Then, however, instead of dropping all of the predictors that do not significantly enhance the prediction, only the predictor variable with the weakest individual correlation with Y is eliminated and the regression equation is recalculated. If there is no significant decrease in the ability to predict Y this predictor variable is eliminated. This process continues until there is a significant effect of dropping a predictor variable.

Simple multiple regression – A form of multiple regression in which all of the predictor variables are assessed simultaneously. All predictors that do not significantly enhance the overall prediction are dropped.

Hierarchical multiple regression – A form of multiple regression in which the researcher specifies the order in which predictor variables are entered into the regression equation. Only those variables that provide a statistically significant improvement in the ability to predict Y are maintained in the equation.

Forward stepwise multiple regression – A form of multiple regression in which predictor variables are entered into the regression equation based upon their individual correlation with Y. To begin, the predictor variable with the strongest individual correlation with Y is entered into the regression equation. Then the variable that accounts for the greatest proportion of the remaining variability is entered. This process continues until there is no longer a statistically significant improvement in the ability to predict Y.

Backward stepwise multiple regression – A form of multiple regression in which all of the predictor variables are assessed simultaneously as in simple multiple regression. Then, however, instead of dropping all of the predictors that do not significantly enhance the prediction, only the predictor variable with the weakest individual *correlation with Y is eliminated and the regression equation is recalculated. If there is no significant decrease in the ability to predict Y this predictor variable is eliminated. This process continues until there is a significant effect of dropping a predictor variable.*

So, which approach to multiple regression should you employ? It has been found that both the forward and backward stepwise multiple regression techniques are likely to be affected by chance factors related to sample selection and thus their results are less stable. As a consequence these approaches are less commonly used. Hierarchical multiple regression is appropriate if the researcher has a theoretical reason for controlling the order in which variables are entered into the regression equation. Otherwise, simple multiple regression is the procedure that is best suited for most researchers.

The first step in conducting a multiple regression analysis is to ensure that the assumptions have been met (these are reviewed later in this appendix). Next, any outliers are identified and either the data from those subjects are omitted or the data are transformed (reviewed in more advanced statistical texts). The analysis is then conducted, presumably with a computer package such as SPSS.

Table D.2 presents the SPSS output of a simple multiple regression analysis. The data come from a study by Norvilitis and Reid (2011) which, in part, examined four possible predictors of academic adjustment in college students (a low score indicated better adjustment). The first section of the output lists the predictors. The four predictors consisted of (1) the hyperactivity component of ADHD, 'adhdhyp', (2) a measure of study skills, 'studytot', (3) the outcome of the 15-item Appreciation of the Liberal Arts scale, 'alastot', and finally (4) the inattention component of ADHD, 'adhdatt'.

The second section of the output, which has the heading 'Model Summary', indicates that the value of the multiple correlation, R, is .731. Recall that with only one predictor, as was reviewed in Chapter 14, we calculate a Pearson correlation. With multiple predictors we calculate a multiple correlation based upon all of the predictor variables. This correlation is symbolized by the capital letter R. Further, in linear regression, which has only one predictor, if we have a statistically significant r we then calculate r^2 in order to determine the proportion of variability in Y explained by the variability in X. Similarly, in a multiple regression analysis we calculate R^2, the proportion of variability in Y accounted for by all of the predictors combined. It is the measure of effect size for a multiple regression. In this example the value of R^2 is .534 (in SPSS this is called R Square). In other words over half of the variability in the academic adjustment scores is accounted for by the four predictor variables. However, it has been found that R^2 slightly overestimates the proportion of variance that has been explained. Consequently, SPSS provides an adjusted R^2 which corrects for this overestimate. In this case the value of the Adjusted R Square is .515. Finally, a measure of variability is given. The smaller the standard error of estimate, the more accurately you can predict Y.

The third section of the output, labeled 'ANOVA', provides a test of the statistical significance of the multiple regression. The ANOVA is testing whether the combined set of predictors explain or account for a significant amount of the variability in our criterion variable, Y. If the ANOVA is not significant there is no point in reviewing the remainder of the output, for there is no significant relationship between your set of predictors and the criterion variable, Y. However, in this case the value of the F ratio is 27.231, which indicates our multiple correlation of .731 has a probability of less than 1 in 1000 (Sig. equals .000) of having occurred by chance if the value was actually zero. As the ANOVA is statistically significant you proceed to the remaining portion of the output.

The fourth section of the output, with the heading 'Coefficients', consists of six columns. The first column includes the labels for each of the predictors in the regression equation as well as the word (Constant). The next column has the heading 'B'. It consists of each of the values for B in the regression equation beginning with the value for the constant (B_0) followed by the regression weights of the predictors (B_1, etc.). The third column provides a measure of variability for each of these values of B. The fourth column, with the heading **Beta**, provides measures of standardized coefficients. These are the values in the regression equation that would be obtained if all of the data had initially been converted to z scores (in other words, standardized), as was reviewed in Chapter 4. Thus Beta is the term used in multiple regression for a regression weight based upon standardized scores (note that beta can also refer to the size of Type II error, which is a different concept). Essentially, the use of standardized scores takes the variability of the predictor measures into account (standardizes them) which permits a more accurate understanding of the effect of each of the predictors. In this case, note that the absolute value of Beta for Adhdhyp is only .022, which is close to zero. We will discuss the meaning of this in a moment.

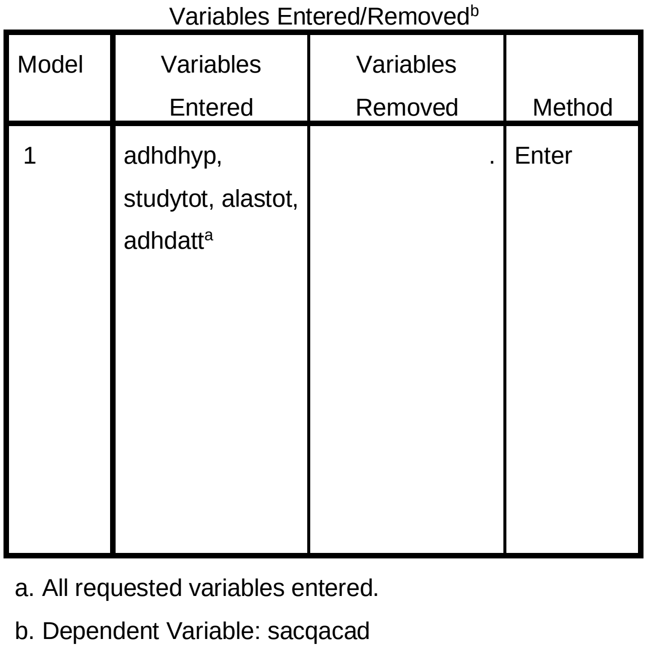
The entries in the fifth column are t values (refer to Chapter 10 for a discussion of the t test). Each predictor is listed in order based upon the absolute value of its t statistic. More specifically, the t statistics provide tests of whether the regression coefficients listed in the second column, with the heading B, are significantly different than 0. In other words, while the ANOVA provided a test of whether the combined set of predictor variables accounted for a significant amount of the variability in the criterion variable, the t tests indicate the significance of each predictor variable. The actual significance levels of these t statistics are provided in the last column, which has the heading 'Sig.'. This column indicates that the B values for Constant, Studytot and Alastot all have a probability of less than 1 in 1000 of having occurred by chance if their actual value for B was zero. The predictor, Adhdatt, has a probability of .05. In other words, the probability is only 1 in 20 of this value of B occurring by chance if the actual value was zero. Thus, two of the four predictors (Studytot and Alastot) have B values statistically different from zero, and another variable, Adhdatt, might be described as being marginally significant. However, the predictor Adhdhyp has a B value of only .098. This B value is associated with a t statistic of .266, and the Sig. column indicates that a B value of this magnitude has a probability of occurring by chance approximately 79 times out of 100 if the actual value of B was zero. This is not statistically significant and thus Adhdhyp, the hyperactivity component of ADHD, was not a significant predictor of academic adjustment in these college students.

Beta – The term used in multiple regression for a regression weight based upon standardized scores.

A more complete output would provide information concerning what is known as **multicollinearity**. Essentially, this concept deals with whether the predictors in the multiple regression equation are highly correlated with each other. If they are, then the results of the multiple regression analysis are likely to vary dramatically between samples. This is an undesirable characteristic. For a detailed discussion you should turn to a more advanced text.

Multicollinearity – In multiple regression a concern that arises when the predictors in the multiple regression equation are highly correlated with each other. If they are, then the results of the multiple regression analysis are likely to vary dramatically between samples. This is an undesirable characteristic.

Table D.2 SPSS Output for a Simple Multiple Regression



Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	adhdhyp, studytot, alastot, adhdatt[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: sacqacad

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.731[a]	.534	.515	9.67358

a. Predictors: (Constant), adhdhyp, studytot, alastot, adhdatt

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10193.061	4	2548.265	27.231	.000[a]
	Residual	8889.929	95	93.578		
	Total	19082.990	99			

a. Predictors: (Constant), adhdhyp, studytot, alastot, adhdatt

b. Dependent Variable: sacqacad

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	119.296	11.405		10.460	.000
	studytot	–.977	.164	–.479	–5.947	.000
	alastot	–.557	.125	–.323	–4.447	.000
	adhdatt	.759	.382	.178	1.989	.050
	adhdhyp	.098	.369	.022	.266	.791

a. Dependent Variable: sacqacad

Reporting The Results Of A Multiple Regression Analysis

Clearly, the SPSS output provides a great deal of information. Table D.3 presents a summary of how the results of this regression analysis would be presented in a paper.

Table D.3 Summary of the Regression Analysis Predicting Academic Adjustment

Predictor Variable	B	SE B	β	t	α
Study skills	–.98	.16	–.48	–5.95	<.001
ALAS	–.56	.13	–.32	–4.45	<.001
Inattention	.76	.38	.18	1.99	.05
Hyperactivity	.10	.37	.02	.27	.79

The multiple regression equation for these data would therefore be:

$$\hat{Y} = (–.98)\ X_1 + (–.56)\ X_2 + (.76)\ X_3 + 119.30$$

This equation can be rewritten as:

$$\hat{Y} = –.98\ X_1 –.56\ X_2 + .76\ X_3 + 119.30$$

where X_1 is the score for Study skills, X_2 is the ALAS score, and X_3 is the Inattention score. Note that the variable Hyperactivity has been dropped as it did not significantly enhance the overall prediction. The value of the Constant, 119.30, comes from the section of Table D.2 with the heading 'Coefficients' and is rounded to two places from 119.296.

Based upon this equation, we predict that our measure of academic adjustment in college students (the criterion or dependent variable, Y) would equal 119.30 if Study skills, ALAS score, and Inattention were all 0. With each increase of 1 point for Study skills, we predict a decrease of 0.98 points for academic adjustment. (It is important to recall that a low score on the measure of academic adjustment in this study signified better adjustment to college.) With each increase of 1 point for the ALAS, we predict a decrease of 0.56 points for academic adjustment. And with each increase of 1 point for Inattention, we predict an increase of 0.76 points for academic adjustment.

Purpose And Limitations Of Multiple Correlation and Multiple Regression

1. *Provides an equation so that the value of Y can be predicted.* The *multiple correlation*, R, provides a measure of the strength of an association between multiple predictor variables (X variables) and a single criterion variable (Y). *Multiple regression* provides an equation for this association which enables Y to be predicted.
2. *Not a measure of cause and effect.* Multiple regression follows the finding of a statistically significant multiple correlation, R. With a study based upon a correlational design the researcher is not justified in coming to a cause-and-effect conclusion due to the lack of

experimental control. Thus, the multiple regression equation allows the prediction of Y from a series of X variables, but does not indicate that the X variables are actually causing changes in the Y variable.

3. *Prediction is limited to the range of the original values.* The multiple regression equation should not be used for values of the predictor variables that are beyond the range of the original data.

Assumptions Of Multiple Correlation and Multiple Regression

1. *Interval or ratio data.* The data are on an interval or a ratio scale of measurement.
2. *Data are associated.* The data are usually multiple measures on the same individual.
3. *Linear relationship.* It is assumed that all of the variables are linearly related.
4. *Significant multiple correlation, R.* A multiple regression analysis will only be undertaken if the multiple correlation, R, has been found to be statistically significant.

Conclusion

Multiple correlation and regression are an extension of the Pearson correlation and linear regression to situations in which there is more than one predictor variable. There are four basic types of multiple regression. For most purposes simple multiple regression is preferable to hierarchical, forward stepwise or backward stepwise multiple regression.

Glossary Of Terms

<u>Backward stepwise multiple regression</u> – *A form of multiple regression in which all of the predictor variables are assessed simultaneously as in simple multiple regression. Then, however, instead of dropping all of the predictors that do not significantly enhance the prediction, only the predictor variable with the weakest individual correlation with Y is eliminated and the regression equation is recalculated. If there is no significant decrease in the ability to predict Y, this predictor variable is eliminated. This process continues until there is a significant effect of dropping a predictor variable.*

<u>Beta</u> – *The term used in multiple regression for a regression weight based upon standardized scores.*

<u>Forward stepwise multiple regression</u> – *A form of multiple regression in which predictor variables are entered into the regression equation based upon their individual correlation with Y. To begin, the predictor variable with the strongest individual correlation with Y is entered into the regression equation. Then the variable that accounts for the greatest proportion of the*

remaining variability is entered. This process continues until there is no longer a statistically significant improvement in the ability to predict Y.

<u>Hierarchical multiple regression</u> – A form of multiple regression in which the researcher specifies the order in which predictor variables are entered into the regression equation. Only those variables that provide a statistically significant improvement in the ability to predict Y are maintained in the equation.

<u>Multicollinearity</u> – In multiple regression a concern that arises when the predictors in the multiple regression equation are highly correlated with each other. If they are, then the results of the multiple regression analysis are likely to vary dramatically between samples. This is an undesirable characteristic.

<u>Simple multiple regression</u> – A form of multiple regression in which all of the predictor variables are assessed simultaneously. All predictors that do not significantly enhance the overall prediction are dropped.

Reference

Norvilitis, J. M. & Reid, H. M. (2011, March). *College success: The relations between appreciation of the liberal arts, symptoms of ADHD and study skills*. Poster presented at the Eastern Psychological Association Convention, Cambridge, MA.

Questions – Appendix D

(Answers are provided in Appendix J.)

1. If the Pearson r is statistically significant then in simple linear regression the _____ will be less than the _____.
   a. mean of Y; mean of X
   b. mean of X; mean of Y
   c. standard error of estimate; standard deviation
   d. standard deviation; standard error of estimate

2. With linear regression there is (are) always _____ predictor(s) while in multiple regression there is (are) always _____ predictor(s).
   a. one; two
   b. one; two or more
   c. two; one
   d. two or more; one

3. In the linear regression equation the Y–intercept is commonly indicated by _____ while in the multiple regression equation the Y–intercept is commonly indicated by _____.
   a. B_1; B_0
   b. a; B_0
   c. B_0; B_1
   d. B_0; a

4. In _____ all of the predictor variables are assessed simultaneously, without any consideration of theoretical importance or prior findings. All predictors that do not significantly enhance the overall prediction are dropped.
   a. simple multiple regression
   b. hierarchical multiple regression
   c. forward stepwise multiple regression
   d. backward stepwise multiple regression

5. This approach to multiple regression is best suited to situations where there is a theoretical basis for the order in which predictor variables are to be considered.
   a. simple multiple regression
   b. hierarchical multiple regression
   c. forward stepwise multiple regression
   d. backward stepwise multiple regression

6. It has been found that this (these) multiple regression technique(s) is (are) likely to be affected by chance factors related to sample selection and thus its (their) results are less stable. As a consequence this (these) approach(es) is (are) less commonly used.
   a. simple multiple regression
   b. hierarchical multiple regression
   c. forward stepwise multiple regression
   d. backward stepwise multiple regression
   e. both 'c' and 'd'

7. In linear regression, if we have a statistically significant r we then calculate _____ in order to determine the proportion of variance in Y explained by the correlation. In multiple regression, if we have a statistically significant R we then calculate _____ in order to determine the proportion of variance in Y explained by the multiple correlation.
   a. R; r
   b. a; B_0
   c. r^2; R^2
   d. B_0; a

8. _____ is the symbol or term used in multiple regression for a regression weight based upon standardized scores.
   a. B_0
   b. B_1
   c. R^2
   d. Beta

9. This concept deals with whether the predictors in the multiple regression equation are highly correlated with each other.
   a. multicollinearity
   b. Beta
   c. hierarchical multiple regression
   d. backward stepwise multiple regression

10. With multiple regression which scale of measurement do we have for the predictors and the criterion?
   a. always interval
   b. either interval or ratio
   c. at least ordinal
   d. any scale of measurement is appropriate.

Appendix E
An Introduction to Power Analysis: Minimum Appropriate Sample Sizes

"A few observations and much reasoning lead to error;
many observations and a little reasoning to truth."
Alexis Carrel

Introduction

This text, like all introductory statistical texts, has emphasized an understanding of Type I error, which is the probability of incorrectly rejecting a true null hypothesis. Type I error is called alpha and has the symbol α (Table E.1 which is a copy of Table 6.3). As was previously discussed, in most experiments the alpha level is set by the researcher at .05, or 1 chance in 20, though another value such as .01 is sometimes chosen.

This appendix will emphasize the importance of Type II error, which is the probability of failing to reject a false null hypothesis. Type II error is called beta (note that beta has a different definition in multiple regression, which is discussed in Appendix D) and has the symbol β (Table E.1). As was noted in Chapter 6, the value of beta is not usually known. Nevertheless, an understanding of beta is important, in part because beta is related to the concept of statistical power, which is defined as 1 – β. Power is thus the probability of correctly rejecting a false null hypothesis, which is what an experimenter is attempting to accomplish when conducting a study (Table E.1). Alternatively, if we assume that an independent variable has had an effect on a dependent variable, then power can be thought of as indicating how well the researcher can detect that this has occurred. An example may be helpful.

Table E.1 Relationship Between Type I and Type II Errors, and Power

		Experimenter's Decision	
		Rejects the Null Hypothesis	Retains the Null Hypothesis
Truth of the Decision (Which is not known.)	If Decision is Incorrect	*Type I error (α)* *Incorrectly rejected a true null*	*Type II error (β)* *Incorrectly retained a false null*
	If Decision is Correct	*Power (1 – β)* *Correctly rejected a false null*	*Correctly retained a true null*

Table E.2 summarizes the outcome of a hypothetical study examining student agreement or opposition to raising college fees to support increased campus services. The null hypothesis is that living on or off campus will not affect students' opinions. If so, there should not be a difference in the distributions of the responses of the two groups. In order to conduct the study, a researcher asks 10 students who live on campus, and 10 who live off campus whether they support raising fees. The data are analyzed with an independent samples chi-square. The chi-square is found to equal 1.6 which, with 1 degree of freedom and $\alpha = .05$, is not statistically significant. Of course, this does not 'prove' that the opinions do not differ, just that there was not sufficient evidence to reject the null hypothesis. But this raises a question, if the population proportions of the two groups of students were in fact different would the researcher have been likely to recognize this? In other words, did the study have enough power to realistically enable the researcher to detect that the proportions differed?

Table E.2 Outcome of Student Opinion Survey on whether Fees should be Raised

	Supports	Opposes
Student lives on campus	7	5
Student lives off campus	3	5

Steps to increase the power of an experiment were reviewed in Chapter 6. One of these steps is to increase the sample size(s) in the study. Thus, when conducting a study a larger sample is, generally speaking, better than a smaller one. However, this advice is of only modest practical value as every experimenter has time and material constraints that will limit how many subjects they can test. What an experimenter needs to know is what the minimum acceptable sample size is for their particular situation, and in Chapter 6 there was no discussion concerning this issue. Further, as this is an introductory text an important consideration in creating examples was that the calculations be kept to a minimum. In order to do so the data were frequently created so that significant outcomes would occur with very small samples. An unfortunate consequence is that a reader may have come to an incorrect view as to the sample size that is likely to be needed in actual research. Determination of how large a sample an experimenter should choose in order to have confidence that they can reasonably expect to reject a false null hypothesis is an example of what is called **power analysis**. The four variables that are involved in power analysis are the desired value for statistical power, the alpha level chosen by the researcher, the effect size, and the sample size.

Power analysis – Detailed examination of the statistical power of a study. The current text emphasizes how this examination can assist the researcher in determining the *minimum sample size that is needed.*

As noted previously, power (1 – β) is the probability of correctly rejecting a false null hypothesis. It has been suggested (Cohen, 1988) that a reasonable level of power for a study would be .80. With this value of power the experimenter would be able to reject a false null hypothesis 80% of the time. Put another way, if the power of a study is .80, then the beta for the study is .20. This is generally seen as a reasonable level for beta. However, a number of published articles have noted that many studies have too little power. In other words, in many studies beta is greater than .20, sometimes much greater, and thus it is unlikely that the null hypothesis could be rejected in these studies even if it was false. What this means is that these experimenters essentially wasted their time conducting their studies! But the damage is not limited to these researchers. Readers of this research may interpret a failure to reject the null hypothesis as evidence that the independent variable in question did not have an effect. If so, they will be less likely to explore this issue in the future and thus a promising direction for research may be overlooked. As many studies are underpowered the cumulative effect can bias the direction or rate of development of an entire field of study (Maxwell, 2004)!

The levels of beta and alpha are linked. Recall that the alpha level is defined as the probability of making a Type I error. It is usually set by the experimenter at .05. Choosing a smaller level of alpha, such as .01, will reduce the probability of making a Type I error but will, of course, simultaneously increase the probability of making a Type II error (refer to Chapter 6 for a review). And, as the probability of Type II error (β) increases, the power of the experiment decreases (remember, power is defined as 1 – β).

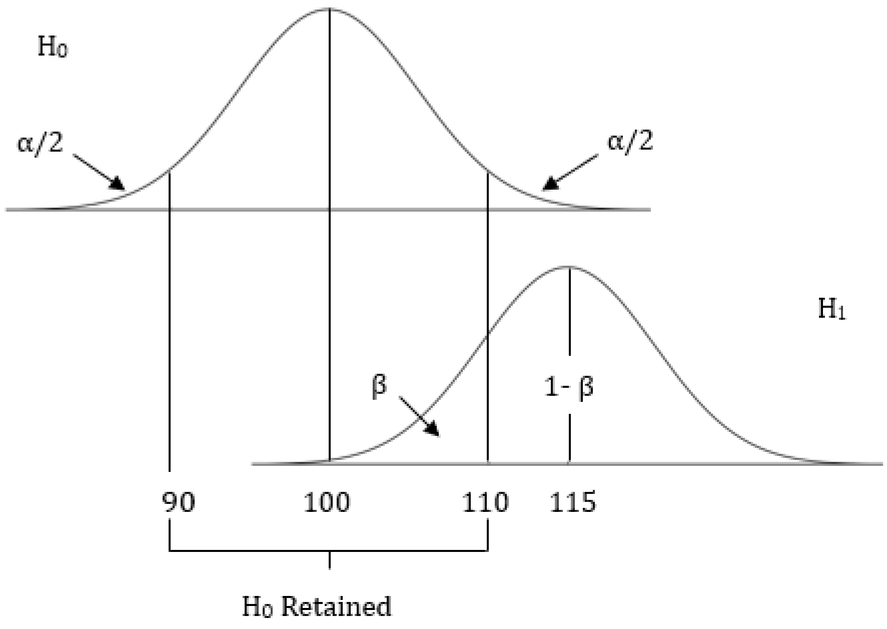
The effect size is also an important component of power analysis. The effect size is a measure of the strength of the independent variable. It can be thought of as the impact that an independent variable has upon a dependent variable. Thus, in a simple two-group study if the effect size is large then the mean of the experimental group is likely to differ substantially from that of the control group. And, of course, all else being equal larger differences are easier to detect than smaller differences.

The final variable in a power analysis is the number of subjects in the study. Intuitively, larger samples provide better estimates of population parameters than do smaller samples. Thus, as a general rule larger samples are better than smaller ones. Further, if the values of the other three variables in a power analysis (desired value of power, the alpha level, and the effect size) are known, then the minimum sample size that is required for a study can also be determined. For instance, it was just noted that a reasonable level of statistical power for a study would be .80. Further, the experimenter sets the value of alpha, usually at .05, so this is a known value. However, determining the effect size is more problematical. A review of previous studies may provide a reasonable estimate.

The relationships of alpha (α), beta (β) and power (1 – β) are illustrated in Figure E.1. Let's assume our null hypothesis is that the population mean is 100. You learned in Chapter 9 that if we were to take numerous samples, all of the same size, from this population, find the mean of each, and graph them, the result would be a sampling distribution of the mean. This distribution is shown in the top portion of Figure E.1. Now assume we conduct an experiment to determine whether our IV had an effect. We collect data from one sample and find its mean. How do we determine whether this sample came from the population as specified by the null hypothesis or from an alternative population with a different mean? Before collecting the data we would have specified the alpha level (usually .05) and also whether we are conducting a one- or two-tailed test. The top portion of Figure E.1 illustrates a two-tailed test as well as the hypothetical critical values for the lower and upper tails, in this case 90 and 110. Thus, if the obtained sample mean is less than 90 or greater than 110 we will reject the null hypothesis and accept that the sample came from a population with a mean different than 100. (This alternative population, with a hypothetical mean of 115, is shown in the bottom portion of Figure E.1.) Of course, our decision to reject the null hypothesis could be incorrect and, by chance, we selected a very unusual sample that actually came from the population specified by the null hypothesis (top portion of Figure E.1). The probability of making this error is determined when we specify the value of alpha (α).

Alternatively, if the sample mean obtained in the experiment was 105 our decision would be to retain the null hypothesis as this value is less than the critical value of 110. Thus we would be concluding that this sample came from the distribution in the top portion of Figure E.1. Of course, once again it is possible that our decision is incorrect. If so, our sample actually came from the population shown in the bottom portion of Figure E.1. The probability of making this error, retaining the null hypothesis when it is incorrect (this is equivalent to not rejecting the null hypothesis when it is incorrect), is beta (β). Then it follows that the probability of correctly rejecting the null hypothesis is 1 – β. This is defined as power and the steps a researcher can take to increase power were discussed in Chapter 6. Clearly, the likelihood of correctly rejecting the null hypothesis will increase if the overlap between the distributions in Figure E.1 is reduced. One way to accomplish this is to increase the sample size.

Figure E.1 Illustration of the Relationships of Alpha (α), Beta (β) and Power (1 – β)



If you have values for power, alpha, and the effect size there are websites that you can use to calculate the needed minimum sample size for your study. Alternatively, you can refer to the following tables to provide an estimate of your needed sample size. These tables are based upon Cohen (1988) and you are encouraged to refer to this reference, or a shorter 'primer' (Cohen, 1992) for a more in-depth discussion of this topic.

The following tables give the minimum sample sizes needed for studies utilizing the chi-square test of independence (Table E.3), one-way between-subjects ANOVA or independent samples t test (Table E.4), and the Pearson correlation (Table E.5). The tables assume that power has been set as .80 and that the experimenter is utilizing a two-tailed test with the alpha level equal to .05. Further, each table includes three estimates for the effect size (small, medium or large). Cohen (1988) provides more extensive tables. However, the current tables provide an easy-to-understand introduction to power analysis and should be useful in determining approximate sample sizes for these commonly utilized statistical procedures.

Table E.3 indicates the approximate total number of subjects needed in a study which utilizes a chi-square test of independence and which has the previously defined values for power and alpha. In this text you were taught to use phi and Cramer's V as measures of effect size with a chi-square test. Cohen's (1988, 1992) tables were based upon the statistic w as the measure of effect size. Following the suggestion of Cohen (1988, 1992), a small effect size would have a w equal to .10, a medium effect size would have a w equal to .30, and a large effect size would have a w equal to .50.

Table E.3 **Determination of the Minimum Number of Subjects Needed for a Chi-square Test of Independence**

	Effect Size		
df	Small	Medium	Large
1	785	87	31
2	963	107	39
3	1090	121	44
4	1194	133	48
5	1293	143	51

It is important to note that in Table E.3 the needed minimum total number of subjects or cases increases as the degrees of freedom increases. This should be intuitively obvious, for it is reasonable that as the number of cells in a design increases so would the necessary size of the total sample. In addition, as the effect size becomes larger the needed sample size decreases dramatically. This is simply an indication that a large difference is easier to detect than a small difference.

Recall that Table E.2 summarized the outcome of a hypothetical study assessing 20 students' opinions. It was concluded that the results did not warrant rejecting the null hypothesis that the opinions did not differ. This was based upon the calculation of a chi-square, with 1 degree of freedom, of 1.6. However, as Table E.3 indicates, even if the effect size was expected to be large the study would have required a minimum sample size of 26 to have adequate power. If a medium effect size was anticipated the minimum sample size would have increased to 87. And the minimum sample size for a small effect would be 785! With a sample size of only 20 the results in Table E.2 are an example of a woefully underpowered study.

Table E.4 indicates the approximate number of subjects needed in each group (level) in a study which utilizes a one-way between-subjects ANOVA or independent samples t test (two-tailed – one df) with the previously defined values for power and alpha. The current text utilized eta^2 as the measure of effect size with ANOVA. The measure of effect size utilized by Cohen (1988, 1992) was the statistic f. Following Cohen's (1988, 1992) suggestion, a small effect size would have an f equal to .10, a medium effect size would have an f equal to .25, and a large effect size would have an f equal to .40.

Table E.4 **Determination of the Minimum Number of Subjects Needed for Each Group in a One-way Between-subjects ANOVA or Independent Samples t Test**

	Effect Size		
df	Small	Medium	Large

1	393	64	26
2	322	52	21
3	274	45	18
4	240	39	16
5	215	35	14

It is important to note that numbers in Table E.4 are for the minimum number of subjects needed in each group (treatment level) of the study. Thus while this number decreases with increasing degrees of freedom, the total number of subjects or cases needed in the entire study still increases as the degrees of freedom increases. In addition, as the effect sizes becomes larger, the number of subjects needed in each group decreases dramatically. Once again, this is simply an indication that it is easier to detect a large difference than a small difference. Finally, it was pointed out in this text that the independent samples t test is closely related to an ANOVA. The number of subjects needed in each group of the independent samples t test is the same as for each group in a one-way between-subjects ANOVA with 1 df_{Bet}.

Table E.5 indicates the approximate number of pairs of observations needed in a study which utilizes a Pearson correlation with the previously defined values for power and alpha. Cohen's (1988, 1992) measure of effect size was based upon the statistic r. This text utilized r^2 as the measure of effect size. Following Cohen's (1988, 1992) suggestion, a small effect size would have an r equal to .10, a medium effect size would have an r equal to .30, and a large effect size would have an r equal to .50.

Table E.5 Determination of the Minimum Number of Pairs of Observations for a Pearson Correlation

Effect Size		
Small	Medium	Large
782	85	29

It is important to note that the entries in Table E.5 are for the number of *pairs* of observations needed in the study. As the effect size becomes larger, the number of pairs of observations needed in the study decreases dramatically. Once again, this is simply an indication that it is easier to detect a large effect than a small effect.

The previous examples are not meant to be comprehensive. A number of sites on the internet are available to determine needed sample sizes. One that is often used is G*Power. Another, very easy-to-use site is Statistical Decision Tree Wizard. An advantage of the later site is that it also provides assistance for choosing the correct statistical procedure to use. However, from even a brief review of the tables provided in this appendix it is evident that substantial sample sizes will frequently be needed in order for a study to have adequate power. This is a critical piece of information when designing your research, or when reviewing the research of others.

Purpose And Limitations Of Using Power Analysis When Choosing Appropriate Sample Sizes

1. *Provides an estimate of the minimum number of subjects that is needed for a study.*
2. *Estimate of the minimum number of subjects needed is dependent upon the values utilized for power, alpha and effect size.* The estimates for power, alpha and effect size are all either based upon convention (power and alpha) or upon previous findings (effect size). While there is a solid logical basis for the conventions that underlie choices for values of power and alpha, there is still the possibility that for a particular study different values would have been appropriate. Further, while the determination of effect size is based upon the previous literature, in many cases this record is likely to be limited. All of these factors could affect the accuracy of the estimate of the minimum number of subjects that is needed.

Assumptions Of Power Analysis When It Is Being Used To Choose Appropriate Sample Sizes

1. *Values of power, alpha and effect size are either known or can be estimated.*

Conclusion

Power analysis provides guidelines for determining the minimum number of subjects that are needed in a study to maintain the Type II error rate at a reasonable level. Without an adequate number of subjects it is likely that a false null hypothesis will not be rejected, a situation which has been found to commonly occur in reviews of published research.

Glossary Of Terms

<u>*Power analysis*</u> *– Detailed examination of the statistical power of a study. The current text emphasizes how this examination can assist the researcher in determining the minimum sample size that is needed.*

References

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. 2nd Ed., Lawrence

Erlbaum Associates, NJ.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*(2), 147–163.

Questions – Appendix E

(Answers are provided in Appendix J.)

1. The probability of incorrectly rejecting a true null hypothesis is the definition of _____.
   a. Type I error
   b. Type II error
   c. Beta
   d. Power
   e. Both ‘b’ and ‘c’

2. The probability of failing to reject a false null hypothesis is the definition of _____.
   a. Type I error
   b. Type II error
   c. Beta
   d. Power
   e. Both ‘b’ and ‘c’

3. The probability of correctly rejecting a false null hypothesis is the definition of _____.
   a. Type I error
   b. Type II error
   c. Beta
   d. Power
   e. Both ‘b’ and ‘c’

4. This statistical concept is defined as 1 – β.
   a. Type I error
   b. Type II error
   c. Beta
   d. Power
   e. Both ‘b’ and ‘c’

5. If the power of a study is .70, then the beta for the study is _____.
   a. .10
   b. .20
   c. .30
   d. .40
   e. cannot be determined

6. Cohen suggested that a reasonable level of power for a study would be _____.
   a. .05
   b. .80
   c. .20
   d. .01
   e. .50

7. As the probability of making a Type I error increases, the probability of making a Type II error _____ and the power of the experiment _____.

a. decreases; increases
b. decreases; decreases
c. increases; decreases
d. increases; increases

8. Assuming that Cohen's recommendation for the size of power has been followed, and alpha has been set at .05, what is the minimum number of subjects that a researcher should plan to include in their chi-square study if there are 4 df and the expected effect size is large?

a. 31
b. 133
c. 48
d. 1194

9. Assuming that Cohen's recommendation for the size of power has been followed, and alpha has been set at .05, what is the minimum number of subjects that a researcher should plan to include in each level of their one-way between-subjects ANOVA study if there are 5 df and the expected effect size is medium?

a. 35
b. 14
c. 215
d. 16

10. Assuming that Cohen's recommendation for the size of power has been followed, and alpha has been set at .05, what is the minimum number of pairs of subjects that a researcher should plan to include in their Pearson r study if the expected effect size is small?

a. 1012
b. 783
c. 85
d. 28

Appendix F
Statistical Symbols Used in this Text
[And Commonly Used Alternatives]

Chapter 1: Math Summary

$\sum X$ "sum each of the scores", it is read as 'sum of X'.

$\sum X^2$ "sum each of the squared scores", it is read as 'sum of X squared'.

$(\sum X)^2$ "the square of the sum of scores", it is read as 'sum of X, quantity squared'

$<$ and $>$ indicate 'less than' and 'greater than', respectively

Chapter 3: Mean and Variability with Interval/Ratio Data

μ	population mean	
M	sample mean	$[\overline{X}]$
σ^2	variance of a population	
σ	standard deviation of a population	

Chapter 4: Variability with Interval/Ratio Data

s^2	variance of a sample
s	standard deviation of a sample
z	z score
α	alpha level

Chapter 7: Chi-square

χ^2	Chi-square
f_o	frequency observed
f_e	frequency expected

Chapter 8: Chi-square

ϕ	Phi

Chapter 9: One Sample z and t tests

σ_M	population standard error	$[\sigma_{\bar{X}}]$
SEM	abbreviation of standard error of the mean	
z_c	critical value of z	
$\leq$	less than or equal to	
t	t score	
s_X	estimate of the population standard deviation based upon sample data	
s_M	estimate of the population standard error based upon sample data	$[\mathbf{s}_{\bar{X}}]$
df	degrees of freedom	

η^2	eta squared
t_c	critical value of t

Chapter 10: Independent Samples t and Dependent Samples t

M_1	mean of sample one
M_2	mean of sample two
$s_{(M_1 - M_2)}$	standard error of the difference between sample means
D	difference between two scores
M_D	mean of a set of difference scores
s_{M_D}	standard error of the mean difference

Chapter 11: One-way Between-subjects ANOVA

F	F ratio in an ANOVA
SS_T	sums of squares total
SS_{Bet}	sums of squares between
SS_W	sums of squares within
df_T	degrees of freedom total
df_{Bet}	degrees of freedom between
df_W	degrees of freedom within
MS_{Bet}	mean square between
MS_W	mean square within
k	number of samples or groups
q	value obtained from the Tukey HSD table

Chapter 12: One-way Within-subjects ANOVA

$SS_{Subjects}$	sums of squares subjects
$SS_{Residual}$	sums of squares residual
$df_{subjects}$	degrees of freedom subjects
$df_{Residual}$	degrees of freedom residual
$MS_{Residual}$	mean square residual

Chapter 13: Two-way Between-subjects ANOVA

F_A	main effect of Factor A
F_B	main effect of Factor B
F_{AXB}	interaction of Factor A and Factor B
df_A	degrees of freedom factor A
df_B	degrees of freedom factor B
df_{AXB}	degrees of freedom for interaction

df_W degrees of freedom within

MS_A mean square factor A

MS_B mean square factor B

MS_{AXB} mean square for interaction

MS_W mean square within

Chapter 14: Pearson Correlation and Regression

r_{XY} Pearson correlation between X and Y

r_{XY}^2 with Pearson correlation, the proportion of variance in one variable that is explained or accounted for by the other variable

M_X mean of the scores of variable X [$\overline{X}$]

M_Y mean of the scores of variable Y [$\overline{Y}$]

$\hat{Y}$ predicted value of Y [Y']

$\sigma_{\hat{Y}}$ standard error of estimate – standard deviation of Y scores around the regression line

$\sigma_{\hat{Y}}^2$ error variance – variance of Y scores around the regression line

b slope of a line, also called the regression weight

a Y intercept of a line

Appendix D: Multiple Regression

B regression weight in multiple regression

R symbol for multiple correlation

R^2 in multiple correlation, the proportion of variance in the Y variable that is explained or accounted for by the set of predictor variables

Appendix G
Definitional Equations and, Where Appropriate, Their Computational Equation Equivalents, In the Order They Were Presented

Descriptive Statistics

Chapter 3: Describing Interval and Ratio Data – I

Median = the value of the score at the $\frac{\mathbf{N+1}}{\mathbf{2}}$ position

Range = highest score – lowest score

Interquartile range = 75th percentile – 25th percentile

$$\text{Semi-interquartile range (SIQR)} = \frac{\textbf{interquartile range}}{\mathbf{2}}$$

$$= \frac{\textbf{75th percentile} - \textbf{25th percentile}}{\mathbf{2}}$$

$$\text{Mean (M)} = \frac{\Sigma X}{n}$$

Population	Sample	Computational Equations
$\sigma^2 = \frac{\Sigma\,(X-\mu)^2}{N}$	$s^2 = \frac{\Sigma\,(X-M)^2}{n-1}$	$s^2 = \frac{\Sigma X^2 - \frac{(\Sigma\,X)^2}{n}}{n-1}$
$\sigma = \sqrt{\frac{\Sigma\,(X-\mu)^2}{N}}$	$s = \sqrt{\frac{\Sigma\,(X-M)^2}{n-1}}$	$s = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma\,X)^2}{n}}{n-1}}$
$\sigma = \sqrt{\frac{SS}{N}}$	$s = \sqrt{\frac{SS}{n-1}}$	
$\sigma = \sqrt{\frac{\Sigma\,x^2}{N}}$	$s = \sqrt{\frac{\Sigma\,x^2}{n-1}}$	

Chapter 4: Describing Interval and Ratio Data

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

$$X = z\sigma + \mu$$

Inferential Statistics

Chapter 7: Goodness-of-fit Chi-square

$$\chi^2 = \Sigma\,\frac{(\text{Frequency observed} - \text{Frequency expected})^2}{\text{Frequency expected}} = \sum \frac{(f_o - f_e)^2}{f_e}$$

$df = c - 1$ where c = the number of categories

Chapter 8: Chi-square Test of Independence

Expected frequency of a cell $= \frac{\text{(Frequency of its row) (Frequency of its column)}}{\text{Total n}}$

df = (Number of rows – 1) (Number of columns – 1)

$\phi = \sqrt{\frac{\chi^2}{n}}$

Cramer’s V $= \sqrt{\frac{\chi^2}{\text{n(df)}}}$

where df = the *smaller* of (r – 1) and (c – 1)

Chapter 9: One Sample z and t tests

$z = \frac{M - \mu}{\sigma_M}$

where $\sigma_M = \frac{\sigma}{\sqrt{n}}$

This equation may be clearer if we substitute σ_X for σ:

$\sigma_M = \frac{\sigma_X}{\sqrt{\mathbf{n}}}$

Confidence interval for z:

$M - z_c\,(\sigma_M) \leq \mu \leq M + z_c\,(\sigma_M)$

where z_c is the critical value for z obtained from the z table (Appendix K, Table 1)

$t = \frac{\mathbf{M} - \mu}{\mathbf{s_M}}$

where $s_M = \frac{\mathbf{s_X}}{\sqrt{\mathbf{n}}}$ and $s_X = \sqrt{\frac{\Sigma(\mathbf{X} - \mathbf{M})^2}{\mathbf{n - 1}}}$

df = n – 1 where n = the number of data points

$\eta^2 = \frac{\mathbf{t^2}}{\mathbf{t^2 + df}}$

Confidence interval for t:

$M - t_c\,(s_M) \leq \mu \leq M + t_c\,(s_M)$

where t_c is the critical value for t obtained from the t table (Appendix K, Table 3b)

Chapter 10: Independent Samples t and Dependent Samples t Tests

<u>Independent Samples t Test</u>

$t = \frac{(\mathbf{M_1} - \mathbf{M_2}) - (\mu_1 - \mu_2)}{\mathbf{s}_{(\mathbf{M_1} - \mathbf{M_2})}}$

where $\mathbf{s}_{(\mathbf{M_1} - \mathbf{M_2})} = \sqrt{\frac{n_1 - 1)\, s_{X_1}^2 + (n_2 - 1)\, s_{X_2}^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

and where df = $n_1 + n_2 - 2$

$\eta^2 = \frac{\mathbf{t^2}}{\mathbf{t^2 + df}}$

Confidence interval:

$[(M_1 - M_2) - t_c\,(\mathbf{s_{(M_1 - M_2)}})] \leq (\mu_1 - \mu_2) \leq [(M_1 - M_2) + t_c\,(\mathbf{s_{(M_1 - M_2)}})]$

where t_c is the critical value for t obtained from the t table (Appendix K, Table 3b)

<u>Dependent Samples t Test</u>

$t = \frac{\mathbf{M_D - \mu_D}}{\mathbf{s_{M_D}}}$

Where the mean difference, $\mathbf{M_D} = \frac{\mathbf{\Sigma D}}{\mathbf{n}}$; the standard error, $\mathbf{s_{M_D}} = \frac{\mathbf{s_D}}{\mathbf{\sqrt{n}}}$; the standard deviation of the differences, $s_D = \sqrt{\frac{\Sigma(\mathbf{D - M_D})^2}{\boldsymbol{n} - \mathbf{1}}}$; and n is equal to the number of *pairs* of scores.

And where df is equal to n – 1

If the null hypothesis is that the difference between population means (μ_D) is zero:

$t = \frac{\mathbf{M_D}}{\mathbf{s_{M_D}}}$

$\eta^2 = \frac{\mathbf{t^2}}{\mathbf{t^2 + df}}$

Confidence interval:

$M_D - t_c\,(\mathbf{s_{M_D}}) \leq \mu_D \leq M_D + t_c\,(\mathbf{s_{M_D}})$

where t_c is the critical value for t obtained from the t table (Appendix K, Table 3b)

Chapter 11: One-way Between-subjects ANOVA

Likelihood of at least one Type I error = $1 - (1 - \alpha)^C$

where c is the number of pairwise comparisons.

Number of pairwise comparisons, $c = \frac{\mathbf{k(k - 1)}}{\mathbf{2}}$

where k is the number of levels of the treatment (number of samples or groups)

$F = \frac{\textbf{Between groups estimate of } \sigma_X^2}{\textbf{Within groups estimate of } \sigma_X^2}$

$SS_T = SS_{Bet} + SS_W$

$SS_T = \Sigma(X - M_G)^2$

where $M_G = \frac{\mathbf{\Sigma X}}{\mathbf{N}}$

M_G is equal to the mean of all of the subjects in all of the samples, and N is the total number of subjects in all the samples or groups

$SS_{Bet} = \Sigma[(M - M_G)^2 n]$

where M is the mean of a sample or group and n is the sample size, the number of subjects in each group or sample

$SS_W = \Sigma[\Sigma (X - M)^2]$

$df_T = df_{Bet} + df_W$

$df_T = N - 1$

$df_{Bet} = k - 1$ where k is the number of groups or treatment levels

$df_W = \Sigma(n - 1)$

$MS_{Bet} = \frac{\mathbf{SS_{Bet}}}{\mathbf{df_{Bet}}}$

$MS_W = \frac{\mathbf{SS_W}}{\mathbf{df_W}}$

$$F = \frac{\mathbf{MS_{Bet}}}{\mathbf{MS_W}} = \frac{\textbf{Estimate of } \sigma_X^2 \textbf{ based upon Treatment + Error}}{\textbf{Estimate of } \sigma_X^2 \textbf{ based only upon Error}}$$

Tukey HSD

Critical value = $q \sqrt{\frac{\mathbf{MS_W}}{\mathbf{n}}}$

where q is found from the q table (Appendix K, Table 5) and n is the number of subjects in *each* sample if the sample size is the same for all of the samples. Or

$$n = \frac{\textbf{Number of means}}{\Sigma \frac{\mathbf{1}}{\textbf{Number of subjects in each sample}}}$$

if the sample size is not the same for all of the samples

$\eta^2 = \frac{\mathbf{SS_{Bet}}}{\mathbf{SS_T}}$

$t^2 = F$

Chapter 12: One-way Within-subjects ANOVA

$SS_T = SS_{Bet} + SS_{Subjects} + SS_{Residual}$

$SS_T = \Sigma(X - M_G)^2$

where $M_G = \frac{\mathbf{\Sigma X}}{\mathbf{N}}$

M_G is equal to the mean of all of the scores in the study, and N is the total number of data points in the study

$SS_{Bet} = \Sigma[(M - M_G)^2 n]$

where n is the number of subjects in the study

$SS_{Residual} = SS_W - SS_{Subjects}$

$SS_W = \Sigma[\Sigma(X - M)^2]$

where M is the mean of a treatment level

$SS_{Subjects} = [\Sigma(\frac{\Sigma \mathbf{X}_{\mathbf{Subject}}}{\mathbf{k}} - \mathbf{M}_{\mathbf{G}})^{\mathbf{2}}]\mathbf{k}$

$= [\Sigma(\mathbf{M}_{\mathbf{Subject}} - \mathbf{M}_{\mathbf{G}})^{\mathbf{2}}]\mathbf{k}$

where k is the number of treatment levels

$df_T = df_{Bet} + df_{Subjects} + df_{Residual}$

$df_T = N - 1$

$df_{Bet} = k - 1$

$df_{Subjects} = n - 1$

$df_{Residual} = (n - 1)(k - 1)$

$MS_{Bet} = \frac{\mathbf{SS}_{\mathbf{Bet}}}{\mathbf{df}_{\mathbf{Bet}}}$

$MS_{Residual} = \frac{\mathbf{SS}_{\mathbf{Residual}}}{\mathbf{df}_{\mathbf{Residual}}}$

$F = \frac{\mathbf{MS}_{\mathbf{Bet}}}{\mathbf{MS}_{\mathbf{Residual}}}$

Number of pairwise comparisons $= \frac{\mathbf{k(k - 1)}}{\mathbf{2}}$

where k is the number of treatment levels

Tukey HSD critical value $= q \sqrt{} \frac{\mathbf{MS}_{\mathbf{Residual}}}{\mathbf{n}}$

where q is found from the q Table (Appendix K, Table 5), $MS_{Residual}$ comes from the ANOVA summary table and n = the number of subjects

$\eta_{\mathbf{P}}^{\mathbf{2}} = \frac{\mathbf{SS}_{\mathbf{Bet}}}{\mathbf{SS}_{\mathbf{T}} - \mathbf{SS}_{\mathbf{Subjects}}}$

Chapter 13: Two-way Between-subjects ANOVA

There are three F ratios:

F_A, the main effect of Factor A

F_B, the main effect of Factor B

F_{AXB}, the interaction of Factor A and Factor B

$df_T = df_A + df_B + df_{AXB} + df_W$

$df_T = N - 1$ where N is the total number of subjects in the study

df_A = number of levels of Factor A – 1

df_B = number of levels of Factor B – 1

$df_{AXB} = df_A \text{ X } df_B$

df_W = N – number of cells

$MS_A = \frac{\mathbf{SS_A}}{\mathbf{df_A}}$

$MS_B = \frac{\mathbf{SS_B}}{\mathbf{df_B}}$

$MS_{AXB} = \frac{\mathbf{SS_{AXB}}}{\mathbf{df_{AXB}}}$

$MS_W = \frac{\mathbf{SS_W}}{\mathbf{df_W}}$

$F_A = \frac{\mathbf{MS_A}}{\mathbf{MS_W}}$

$F_B = \frac{\mathbf{MS_B}}{\mathbf{MS_W}}$

$F_{AXB} = \frac{\mathbf{MS_{AXB}}}{\mathbf{MS_W}}$

Number of pairwise comparisons $= \frac{\mathbf{k(k - 1)}}{\mathbf{2}}$

where k is the number of means being compared

Tukey HSD critical value for a main effect $= q \sqrt{} \frac{\mathbf{MS_W}}{\mathbf{n}}$

where q is found from the q table (Appendix K, Table 5), MS_W comes from the ANOVA summary table and n = the number of scores for *each* mean

Tukey HSD critical value for an interaction $= q_i \sqrt{} \frac{\mathbf{MS_W}}{\mathbf{n}}$

Where q_i is derived from q (refer to a more advanced text)

Eta squared (η^2) is calculated for each F ratio that was found to be significant:

η^2 for Factor A $= \frac{\mathbf{SS_A}}{\mathbf{SS_T}}$

η^2 for Factor B $= \frac{\mathbf{SS_B}}{\mathbf{SS_T}}$

η^2 for the Interaction $= \frac{\mathbf{SS_{AXB}}}{\mathbf{SS_T}}$

Alternatively, partial eta squared ($\eta^{\mathbf{2}}_{\mathbf{p}}$) can be calculated for each F ratio that was found to be significant:

$\eta^{\mathbf{2}}_{\mathbf{p}}$ for Factor A $= \frac{\mathbf{SS_A}}{\mathbf{SS_T - SS_B - SS_{AXB}}}$

$\eta^{\mathbf{2}}_{\mathbf{p}}$ for Factor B $= \frac{\mathbf{SS_B}}{\mathbf{SS_T - SS_A - SS_{AXB}}}$

$\eta^{\mathbf{2}}_{\mathbf{p}}$ for the interaction $= \frac{\mathbf{SS_{AXB}}}{\mathbf{SS_T - SS_A - SS_B}}$

Chapter 14: Pearson Correlation and Regression

Pearson Correlation

$\mathrm{cov}_{xy} = \frac{\sum(X - M_X)(Y - M_Y)}{n - 1}$

where n is the number of *pairs* of scores

$r_{XY} = r = \frac{\mathrm{cov}_{XY}}{s_X s_Y}$

df for the Pearson r = n – 2, where n is the number of *pairs* of scores

Coefficient of determination = r^2 which is the proportion of variance in one variable that is explained or accounted for by the other variable

Coefficient of nondetermination = 1 – r^2

Regression

$\sigma_{\hat{Y}} = \sigma_Y \sqrt{}(1 - r^2)$

$r^2 = \frac{\sigma^2_Y - \sigma^2_{\hat{Y}}}{\sigma^2_Y}$

$\hat{Y} = bX + a$

where $b = \frac{\text{Change in Y}}{\text{Change in X}}$

$b = r\,(\frac{s_Y}{s_X})$

$a = M_Y - bM_X$

Definitional Equations	Computational Equations
$b = r\,(\frac{\sigma_Y}{\sigma_X})$	$b = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2}$

Appendix A: Kruskal-Wallis H Test

$$H = [\frac{12}{N(N+1)}][\Sigma(\frac{T^2}{n})] - 3(N+1)$$

where N = the total number of subjects, T = the total of the ranks for a sample, n = the sample size, and df = number of samples – 1

number of possible pairwise comparisons $= \frac{k(k-1)}{2}$

where k = the number of samples

Eta squared $(\eta^2) = \frac{H}{N-1}$

Appendix B: Phi Correlation

$r_\phi = (ad - bc) / \sqrt{}[(a+b)(c+d)(a+c)(b+d)]$

Coefficient of determination = r_ϕ^2

Coefficient of nondetermination = $1 - r_\phi^2$

Appendix C: Spearman Correlation

$r_s = 1 - \frac{6\Sigma D^2}{n(n^2-1)}$

where D is the difference between a pair of ranks and n is the number of *pairs* of data

The Spearman r table (Appendix K, Table 7) is not based upon degrees of freedom but rather is based directly upon n, the number of pairs of data

Coefficient of determination = r_S^2

Coefficient of nondetermination = $1 - r_S^2$

Appendix D: Multiple Regression

The equation for multiple regression with two predictors is:

$\hat{Y} = B_1X_1 + B_2X_2 + B_0$

Where $\hat{Y}$ is the predicted value of the criterion variable; B_1 is the regression weight associated with the first predictor variable, X_1; B_2 is the regression weight associated with the second predictor variable, X_2; and B_0 is a constant, the value of the Y intercept

The equation for multiple regression with three predictors could be written as:

$\hat{Y} = B_1X_1 + B_2X_2 + B_3X_3 + B_0$

Where each term is defined as before and where B_3 is the regression weight associated with the third predictor variable, X_3

Appendix H
Inferential Statistical Procedures and Their Measures of Effect Size Which were Discussed in the Text

Statistical Procedure	Measure of Effect Size

For Difference Designs (Presented In Order Of Coverage)

Goodness-of-fit Chi-square	None
Chi-square test of independence	Phi or Cramer's V
One-sample t test	Eta squared
Independent samples t test	Eta squared
Dependent samples t test	Eta squared
One-way Between-subjects ANOVA	Eta squared
One-way Within-subjects ANOVA	Partial eta squared
Two-way Between-subjects ANOVA	Eta squared or Partial eta squared
Kruskal-Wallis H test (Appendix A)	Eta squared

For Association Designs (Presented In Order Of Coverage)

Chi-square test of independence	Phi or Cramer's V
Pearson correlation	Pearson r squared
Phi correlation (Appendix B)	Phi r squared
Spearman correlation (Appendix C)	Spearman r squared

Appendix I – 1
Glossary of Terms
With Chapter/Appendix Each is First Introduced

Absolute value – The magnitude of a number irrespective of whether it is positive or negative. Chap 1

Alpha – Another term for Type I error. Its symbol is α. Chap 6

Alpha level – Criterion set for rejecting the null hypothesis. This is usually .05. Chap 6

Alternative hypothesis (H_1) – When used with a difference design, the statement that the treatment does have an effect. Chap 6

Analysis of variance (ANOVA) – A set of flexible, closely related, inferential procedures for comparing sample means by examining variances. Chap 11

Area of rejection – Area of the distribution equal to the alpha level. It is also called the Critical Region. Chap 7

Association design – A research procedure designed to determine whether an association observed in a sample is likely to generalize to the population. Chap 6

Backward stepwise multiple regression – A form of multiple regression in which all of the predictor variables are assessed simultaneously as in simple multiple regression. Then, however, instead of dropping all of the predictors that do not significantly enhance the prediction, only the predictor variable with the weakest individual correlation with Y is eliminated and the regression equation is recalculated. If there is no significant decrease in the ability to predict Y, this predictor variable is eliminated. This process continues until there is a significant effect of dropping a predictor variable. App D

Bar graph – A graph in which the frequency of each category or class of observation is indicated by the length of its associated bar. Chap 2

Bell-shaped curve – A symmetrical distribution in which the highest frequency scores are located near the middle and the frequency drops the farther a score is from the middle. Chap 3

Beta – Another term for Type II error. Its symbol is β. Chap 6

also

– The term used in multiple regression for a regression weight based upon standardized scores. App D

Between-subjects design – With an ANOVA, those designs in which each subject experiences only a single level of a factor. Chap 11

Biased estimator – An estimator that does not accurately predict what it is intended to because of systematic error. Chap 9

Bimodal – A descriptive term for a distribution that has two modes. Chap 2

'Blind' study – A study in which the data are collected in such a way that the subject's assignment to the control or experimental condition is not known. There are several variations of 'blind' procedures. They are all employed to reduce bias. Chap 15

Bonferroni method – A procedure to control the Type I error rate when making numerous comparisons. In this procedure the alpha level that the experimenter has set is divided by the number of comparisons. Chap 8

Box and whiskers plot – Another name for a boxplot. Chap 3

Boxplot – A summary of a distribution which includes the median, a central box with the 25th and 75th percentiles as limits, and the range. Another name for a boxplot is a box and whiskers plot. Chap 3

Carryover effect – A treatment or intervention at one point in time may affect or carry over to another point in time. Chap 10

Cause-and-effect conclusion – Decision that the change in the value of the independent variable resulted in a change in the value of the dependent variable. This is justified with a well-conducted, true experiment. Chap 6

Ceiling effect – When the scores are predominately at the high end of the range of possible outcomes. App A

Cell – A particular combination of treatment levels in a Factorial ANOVA. Chap 13

Central limit theorem –

–With increasing sample sizes, the shape of the distribution of sample means (sampling distribution of the mean) rapidly approximates the normal distribution irrespective of the shape of the population from which it is drawn.

–The mean of the distribution of sample means (M_G) is an unbiased estimator of the population mean.

–And the standard deviation of the distribution of sample means (σ_M) will equal $\sigma_X / \sqrt{n}$. Chap 9

Chi-square test of association – Another name for the chi-square test of independence. Chap 8

Chi-square test of independence – An inferential procedure for analyzing whether the pattern of observed frequencies differs among the groups. Chap 8

Coefficient of determination – The square of the correlation. It indicates the proportion of variability in one variable that is explained or accounted for by the variability in the other variable. Chap 14

Coefficient of nondetermination – The proportion of the variability of one variable not explained or accounted for by the variability of the other variable. For the Pearson r, it is equal to $1 - r^2$. Chap 14

Computational equations – Equations developed to aid in statistical calculations. They were useful with large data sets, but now researchers would employ computer software packages instead. Chap 5

Confidence interval – The range of values that has a known probability of including the population parameter, usually the mean. Chap 9

Confirmation bias – Selecting only evidence that supports, or confirms, one's pre-existing beliefs. Chap 15

Confounded comparison – Comparison of two cell means which involves two factors that are changing. The comparison cannot be interpreted. Chap 13

Continuous variable – A variable that can be of any magnitude, though it might be limited to a particular range. Chap 2

Control group – In a between-groups design, the group of subjects that does not receive the treatment. Chap 6

Correlation – A measure of the degree of association among variables. A correlation indicates whether a variable changes in a predicable manner as another variable changes. Chap 14

Correlation coefficient – A single number that indicates the degree to which two variables are related. Chap 14

Correlational study – A study in which the researcher does not randomly assign the subjects and does not manipulate the value of a variable. As a result, at the conclusion of the study the researcher has little confidence that there is a cause-and-effect relationship between the variables. Chap 6

Counterbalancing – A method used to control for carryover effects. In counterbalancing, the order of the treatments or interventions is balanced so that an equal number of subjects will experience each order of presentation. Chap 10

Covariance – A statistical measure indicating the extent to which two variables vary together. Chap 14

Covary – If knowledge of how one variable changes assists you in predicting the value of another variable, the two variables are said to covary. Chap 14

Cramer's V – Measure of effect size for chi square tests of independence larger than 2 X 2. Chap 8

Criterion variable (Y) in regression – The variable (Y) whose value is being predicted by the predictor variable (X). Chap 14

Critical region – Area of the distribution equal to the alpha level. It is also called the Area of Rejection. Chap 7

Critical value – A value for a statistical test which is used to determine whether to reject or retain the null hypothesis. Chap 7

Data (plural of datum) – Observations or factual information, often in the form of numbers. Chap 1

Data view – SPSS window in which the data are displayed. Chap 5

Deduction – A method of thinking in which conclusions are logically derived from general statements that are assumed to be true. Chap 15

Degrees of freedom (df) – The number of outcomes out of the total that are free to vary. Chap 7

Dependent samples t test – An inferential procedure for comparing two sample means based upon repeated measures of the same subjects, or measures from pairs of subjects who are related in some way. Chap 10

Dependent variable (DV) – In an experiment, the variable whose value is not directly controlled by the researcher. Its value may be changed by the independent variable (IV). Chap 6

Dependent variable (Y) in regression – Another name for the criterion variable. Chap 14

Descriptive statistics – Techniques that are used to summarize data. These procedures lead to a better understanding of the data. Chap 1

Deviation – The difference between a score and its mean. Thus, with population data the deviation equals $X - \mu$*. The symbol for a deviation is x. Chap 3*

Difference design – A research procedure designed to determine whether a difference observed between samples is likely to generalize to the populations. Chap 6

Difference score (D) – The difference between two measurements from the same individual (repeated measures design) or two measurements from pairs of matched subjects (matched subjects design). Chap 10

Discrete variable – A variable that can only have particular values. Chap 2

Effect size – A measure of how 'strong' a statistically significant outcome is. Chap 8

Empiricism – A method for finding truth that emphasizes the importance of observation. Chap 15

Error – An outcome due to chance, Chap 9, or with ANOVA, the variability not due to treatment. Chap 11

Error in an ANOVA – 'Pre–existing subject differences' + 'residual error'. Chap 12

Error variance ($\sigma_{\hat{Y}}^2$) *– The variance of Y scores around the regression line. Chap 14*

Expected frequencies – With nominal data, the outcome that would be expected if the null hypothesis were true. Chap 7

Experimental group – In a between-groups design, the group of subjects that does receive the treatment. Chap 6

Experimentwise error rate – The likelihood of making at least one Type I error with any of the experiment's comparisons. Chap 11

Eta squared (η^2) *– A commonly used measure of effect size that indicates the percentage of variation in the dependent variable that is explained or accounted for by the independent variable. Chap 9*

Factor – With an ANOVA, the term 'Factor' is often used instead of independent variable. Chap 11

With factor analysis, one of a smaller number of underlying variables derived from analysis of the larger set of initial variables. Chap 15

Factor analysis – Statistical procedure that groups the initial variables into a smaller set of underlying variables called factors. Chap 15

Factorial ANOVA – An ANOVA with more than one factor. Chap 11

First quartile – The value of the score at the 25th percentile in a distribution. Chap 3

Fisher exact test – An alternative to the 2 X 2 chi square test of independence that is used when there is a particularly small data set. Chap 15

Forward stepwise multiple regression – A form of multiple regression in which predictor variables are entered into the regression equation based upon their individual correlation with Y. To begin, the predictor variable with the strongest individual correlation with Y is entered into the regression equation. Then the variable that accounts for the greatest proportion of the remaining variability is entered. This process continues until there is no longer a statistically significant improvement in the ability to predict Y. App D

Frequency distribution – A listing of the different values or categories of the observations along with the frequency with which each occurred. Chap 2

Frequency polygon – A graphic presentation for use with interval or ratio data. It is similar to a histogram except that the frequency is indicated by the height of a point rather than the height of a bar. The points are connected by straight lines. Chap 3

Gambler's fallacy – The incorrect assumption that if an event has not occurred recently, then the probability of it occurring in the future increases. Chap 8

Goodness-of-fit chi-square test – An inferential procedure that tests whether observed frequencies differ from expected frequencies. Chap 7

Grand mean (M_G) *– The mean of the sample means. In some statistical procedures it is defined as the mean of all of the scores. Chap 9*

Hierarchical multiple regression – A form of multiple regression in which the researcher specifies the order in which predictor variables are entered into the regression equation. Only those variables that provide a statistically significant improvement in the ability to predict Y are maintained in the equation. App D

Histogram – A graph used with interval/ratio data. As with the bar graph, frequencies are indicated by the length of the associated bars. However, as the data are continuous in a histogram the bars are positioned side-by-side. Chap 3

Hypothesis – A scientifically-based statement about some condition in the environment or population. Chap 6

Hypothesis testing – Statistically analyzing data to evaluate whether the null hypothesis should be *retained or rejected. Chap 6*

Independent – Two events, samples or variables are independent if knowing the outcome of one does not enhance our prediction of the other. Chap 7

Independent samples t test – An inferential procedure for comparing two means from unrelated samples. Chap 10

Independent variable (IV) – In an experiment, the variable the experimenter manipulates or directly controls. Chap 6

Induction – A method of thinking in which conclusions are derived from generalizations based upon limited statements or observations that are assumed to be true. Induction is fundamental to science, as observations are used to develop general laws of nature. Chap 15

Inferential statistics – Techniques that are used in making decisions based upon data. Chap 1

Inflection point – A point on a graph where the curvature changes from concave to convex or from convex to concave. Chap 3

Interaction – A change in the dependent variable that is due to the presence of a particular combination of independent variables. Chap 13

Interquartile range (IQR) – A measure of variability based upon the median that includes the middle 50% of the data. It is the range of values in a distribution between the 25th and 75th percentiles. Chap 3

Interval scale of measurement – A measurement scale in which the magnitude of the difference between numbers is meaningful, and thus addition and subtraction are possible. However, there is no true zero and thus multiplication and division are not meaningful. Chap 2

Intrinsic plausibility – Decision-making process in which the alternative that seems most reasonable is accepted as being true. Chap 6

Kruskal-Wallis H test – An inferential procedure that is analogous to the one-way between-subjects ANOVA except that it is used with ordinal data. App A

Law of large numbers – The larger the sample size, the better the estimate of population parameters such as μ. Chap 9

Leaf – The last digit(s) of a score. With a stem-and-leaf display each leaf is paired with the appropriate stem value and the leaves are listed in ascending order in each row of the display. Chap 3

Level – With an ANOVA, the number of values of an independent variable. Chap 11

Levene's test of equality of variances – Procedure used with SPSS to test the assumption that samples are drawn from populations which have equal variances. Chap 10

Longitudinal study – A study in which subjects are measured repeatedly across time. A repeated-measures design is a type of longitudinal study. Chap 10

Main effect – With a factorial ANOVA, another term for an independent variable or factor. Chap 13

Manipulate – The researcher determines which condition of the independent variable each subject receives. Chap 6

MANOVA – An extension of ANOVA in which there is more than one dependent variable. Chap 15

Matched subjects design – A research design in which equivalent subjects are paired and then one of the subjects is randomly assigned to each group. Chap 10

Mauchly's test of sphericity – Statistical procedure utilized with SPSS to test the assumption of sphericity for a one-way within-subjects ANOVA. Chap 12

Mean – A measure of central tendency for use with interval or ratio data. It is what is commonly called an average. The mean is the sum of the scores divided by the total number of scores. Chap 1 and 3

Mean square (MS) – In an ANOVA, an estimate of the population variance (σ_X^2). Chap 11

Mean square between (MS_{Bet}) – The estimate of the population variance (σ_X^2) based upon the variability between the sample means. More specifically, it is obtained from the deviations of the sample means from the grand mean. Chap 11

Mean square within (MS_W) – The estimate of the population variance (σ_X^2) based upon the variability within each of the samples. More specifically, it is obtained by pooling the variances of the scores within each of the samples. Chap 11

Measure of central tendency – A single number that is chosen to best summarize an entire set of numbers. Chap 2

Median – A measure of central tendency. It is the mid-most score in a distribution. In other words, the median splits a distribution in half, with just as many scores above it as below it. It is at the 50th percentile. Chap 2

Mixed ANOVA – Factorial ANOVA in which there are both between-subjects and within-subjects factors. Chap 15

Mode – A measure of central tendency. It is the most common category or score. Chap 2

Multicollinearity – In multiple regression a concern that arises when the predictors in the multiple regression equation are highly correlated with each other. If they are, then the results of the multiple regression analysis are likely to vary dramatically between samples. This is an undesirable characteristic. App D

Multiple correlation (R) – The association between one criterion variable and a combination of two or more predictor variables. Chap 14

Multiple linear regression – A procedure in which several variables (Xs) are used to predict the value of another variable (Y). Chap 14

Negative correlation – A relationship between two variables in which as one variable increases in value, the other variable decreases in value. Also, as one variable decreases in value, the *other increases in value. Chap 14*

Negatively skewed – A nonsymmetrical distribution in which the tail pointing to the left is larger than the tail pointing to the right. Chap 3

Nominal scale of measurement – A measurement scale in which numbers serve as names of categories. In this level of measurement, the magnitude of the number is arbitrary. Chap 2

Nonparametric procedure – Statistical procedure that does not make assumptions about the population's parameters and does not assume that the population is normally distributed. Chap 7

Normal distribution – A specific, bell-shaped distribution. Many statistical procedures assume that the data are distributed normally. Chap 3

Null hypothesis (H_0) – When used with a difference design, the statement that the treatment does not have an effect. Chap 6

Observed frequencies – With nominal data, the actual data that were collected. Chap 7

One-sample t test – An inferential procedure for comparing a sample mean with a population mean when the population standard deviation is not known. Chap 9

One-sample z test – An inferential procedure for comparing a sample mean with a population mean when the population standard deviation is known. Chap 9

One-tailed or directional test – An analysis in which the null hypothesis will only be rejected if an extreme outcome occurs in the predicted direction. In such a test, the single area of rejection is equal to alpha and it is located in one tail of the sampling distribution. Chap 9

One-way between-subjects ANOVA – An inferential procedure for comparing two or more means from independent samples when there is one independent variable. Chap 11

One-way within-subjects ANOVA – An inferential procedure for comparing two or more means from related samples when there is one independent variable. Chap 12

Ordinal scale of measurement – A measurement scale in which the magnitude of the numbers indicates the order in which events occurred. In this level of measurement, the magnitude of the number is meaningful. Chap 2

p-value – The probability of an outcome, or a more extreme outcome, occurring by chance assuming the null hypothesis is correct. To be statistically significant, the p-value must be less than the alpha level, which is usually .05. Chap 8

Pairwise comparison – Comparison between two sample means. Chap 11

Pairwise error rate – The likelihood of making a Type I error for a single comparison between sample means. This is equal to α, which is usually .05 or .01. Chap 11

Parameter – A measure of a characteristic of a population, such as its mean or its variance. Chap 3

Parametric procedure – Statistical procedure that does make assumptions about the population's parameters and does assume that the population is normally distributed. Chap 7

Partial correlation – A procedure in which the effect of a variable that is not of interest is removed. Chap 14

Partial eta squared (η^2_p) – Measure of effect size calculated by SPSS for a within subjects ANOVA. Chap 12

Percentile rank – The percentage of the data at or below a category or score. Chap 2

Phi – Measure of effect size for the 2 X 2 chi square tests of independence. Chap 8

Phi correlation (r_ϕ) – A form of Pearson correlation used with nominal data when both variables are dichotomous. App B

Pie chart – A presentation of categorical data in which the area of a slice of a circle is indicative of the relative frequency with which the category occurs. Chap 2

Population – The entire group that is of interest. Chap 3

Positive correlation – A relationship between two variables in which as one variable increases in value, so does the other variable. Also, as one variable decreases in value, so does the other. Chap 14

Positively skewed – A nonsymmetrical distribution in which the tail pointing to the right is larger than the tail pointing to the left. Chap 3

Post hoc comparisons – Statistical procedures utilized following an initial, overall test of significance in order to identify the specific conditions (samples) that differ. Chap 8

Power – The probability of correctly rejecting a false null hypothesis. This probability is 1 – β. Chap 6

Power analysis – Detailed examination of the statistical power of a study. The current book emphasizes how this examination can assist the researcher in determining the minimum sample size that is needed. App E

Predictor variable (X) in regression – The variable (X) that is used to predict the value of the dependent or criterion variable (Y). Chap 14

Preexisting subject differences – Relatively stable subject characteristics. These differences between subjects are a form of error in an ANOVA. The variability due to these differences is removed in a one-way within-subjects ANOVA. Chap 12

Quartile – A range of values that includes one fourth, or a quarter, of the scores. Chap 3

Quasi-experiment – An experiment in which some characteristic of a true experiment is missing. Most commonly, the researcher manipulates the value of the independent variable but does not randomly assign the subjects. As a result, at the conclusion of the study the researcher has less confidence in concluding that there is a cause-and-effect relationship between the independent and dependent variables than would be the case with a true experiment. Chap 6

Random sample – A sample in which every member of the population has an equal chance of being chosen. Chap 6

Range – A measure of variability for ordinal data. It is obtained by subtracting the lowest rank from the highest rank. Chap 2

Also, a measure of variability for interval or ratio data. It is commonly defined as the value which is obtained when the lowest score is subtracted from the highest score. Chap 3

Ratio scale of measurement – A measurement scale in which the magnitude of the difference between numbers is meaningful, and there is a true zero. Thus, multiplication and division as well as addition and subtraction are meaningful. Chap 2

Rationalism – A method for finding truth that emphasizes logical thinking rather than observation. Chap 15

Raw score – Your data as they are originally measured, before any transformation. Chap 4

Real limits – With interval or ratio data, the actual limits used in assigning a measurement. These are halfway between adjacent scores, and are called the upper and lower real limits. Chap 3

Region of rejection – Area of the distribution equal to the alpha level. It is also called the Critical Region. Chap 6

Regression – Procedure researchers use to develop an equation that permits the prediction of one variable of a correlation if the value of the other variable is known. Chap 14

Regression line – With linear regression, a straight line indicating the value of Y that is predicted to occur for each value of X. The symbol for the predicted value of Y is $\hat{Y}$. Chap 14

Regression weight – Another term for the slope of the regression line. Chap 14

Relative frequency – The frequency of a category divided by the total frequency. Chap 2

Repeated measures design – A research design in which each subject is tested more than once. Chap 10

Residual error – Changeable subject characteristics. These differences between subjects are a form of error in an ANOVA. The variability due to these differences is not removed in a one-way within-subjects ANOVA. Chap 12

Restriction of the range – Reducing the range of values for a variable will reduce the size of the correlation. Chap 14

Rho (ρ) – Symbol used for the population correlation. Chap 14

Sample – A subset of a population. Chap 3

Sampling distribution of the mean – A theoretical probability distribution of sample means. The samples are all of the same size and are randomly selected from the same population. Chap 9

Sample of convenience – A sample that is chosen because it is easily available rather than because it is optimal. Chap 15

Scientific method – An approach to understanding that emphasizes rigorous logic, but also that *careful observation is the ultimate authority for determining truth. It is a self-correcting approach that limits bias. Chap 6*

Second quartile – The value of the score at the 50th percentile in a distribution. It is the median. Chap 3

Semi-interquartile range (SIQR) – A commonly used measure of variability, particularly for skewed data. It is equal to half of the interquartile range. Chap 3

Significance level – Another term for alpha level, the criterion set for rejecting the null hypothesis. This is usually .05. Chap 6

Significant – In statistics, the conclusion that an outcome is unlikely to have occurred by chance. Chap 7

Simple linear regression – Procedure used to determine the equation for the regression line. Chap 14

Simple multiple regression – A form of multiple regression in which all of the predictor variables are assessed simultaneously. All predictors that do not significantly enhance the overall prediction are dropped. App D

Skewed – A distribution in which one tail is larger than the other. As a result, the distribution is not symmetrical. Chap 3

Slope of the line – One of the two determinants of the equation for a straight line. It is the ratio of the change in the Y variable divided by the change in the X variable. It has the symbol 'b' in the equation $Y = bX + a$*. Chap 14*

Spearman correlation (r_S*) – A form of Pearson correlation used when the two variables are measured at the ordinal level. App C*

Sphericity – Assumption of a within-subjects ANOVA that the variances of the sets of difference scores between treatment levels are equal. In a repeated measures ANOVA these differences would be based upon pairs of scores from each subject. Chap 12

SPSS – A powerful, commonly-used statistical computer package. The letters 'SPSS' originally were an abbreviation for 'statistical package for the social sciences'. Chap 5

Standard deviation – A measure of variability; the expected deviation of a score from its mean. It is defined as the square root of the variance. The symbol for the population standard deviation is σ. Chap 3

Standard error of estimate ($\sigma_{\hat{Y}}$ *) – The standard deviation of Y scores around the regression line. Chap 14*

Standard error of the mean (SEM) – The standard deviation of the sampling distribution of means. Chap 9

Standard error of the difference between sample means ($\boldsymbol{s}_{(M_1 - M_2)}$*) – The standard deviation of the sampling distribution of the difference between sample means. Chap 10*

Standard error of the mean difference (s_{M_D}) – *The standard deviation of the means of difference scores. More precisely, the standard deviation of the sampling distribution of the means of difference scores. Chap 10*

Standard score – A measure indicating whether a score is above or below the mean as well as how many standard deviations it is from the mean. Also called a z score. Chap 4

Statistic – A measure of a characteristic of a sample, such as its mean or its variance. Chap 3

Stem – With a stem-and-leaf display, a list of the different values of the data once the last digit(s) of each score is removed. Chap 3

Stem-and-leaf display – A commonly used summary of interval or ratio data in which each original score is separated into two parts, a stem and a leaf. Chap 3

Sum of the squared deviations – For a population, it is equal to $\Sigma(X - \mu)^2$ *or* Σx^2. *It is often abbreviated as 'sum of squares' which is shortened even further to SS. Chap 3*

Sum of squares between groups (SS_{Bet}) – *The sum of the squared deviations of each treatment mean from the grand mean. Chap 11*

Sum of squares residual ($SS_{Residual}$) – *In a one-way within-subjects ANOVA, the SS due to residual error. Chap 12*

Sum of squares subjects ($SS_{Subjects}$) – *In a one-way within-subjects ANOVA, the SS due to preexisting subject differences. Chap 12*

Sum of squares total (SS_T) – *The sum of the squared deviations from the mean for all of the scores. Chap 11*

Sum of squares within groups (SS_W) – *The sum across all conditions, of the sum of the squared deviations of each score from its treatment mean. Chap 11*

Symmetrical distribution – A distribution in which the right half is the mirror image of the left half. In such a distribution, there is a high score corresponding to each low score. Chap 3

Third quartile – The value of the score at the 75th percentile in a distribution. Chap 3

Treatment – With ANOVA, another term for the independent variable. Chap 11

Trend analysis – A statistical technique that attempts to define patterns in data. Chap 12

True experiment – An experiment in which the researcher randomly assigns the subjects and also manipulates the value of the independent variable. As a result, at the conclusion of the study the researcher is justified in reaching a cause-and-effect conclusion concerning the relationship between the independent and dependent variables. Chap 6

Tukey HSD – A popular post hoc test used with ANOVAs. Chap 11

Two-way between-subjects ANOVA – An inferential procedure for comparing means from independent samples when there are two independent variables. Chap 13

Two-tailed or nondirectional test – An analysis in which the null hypothesis will be rejected if an *extreme outcome occurs in either direction. In such a test, the area of rejection is divided into two parts, each equal to α / 2. Chap 9*

<u>Type I error</u> – The probability of rejecting the null hypothesis when it is in fact true. This probability is equal to alpha, α, which is usually 5%. Chap 6

<u>Type II error</u> – The probability of retaining the null hypothesis when it is in fact false. This probability is equal to beta, β. The probability of β is usually not known. Chap 6

<u>Unconfounded comparison</u> – Comparison of two <u>cell</u> means which involves only one factor that is changing . The comparison can be interpreted. Chap 13

<u>Unimodal</u> – A descriptive term for a distribution that has one mode. Chap 2

<u>Unstable</u> – A term used to describe a measure, such as of central tendency, that can vary significantly with only a few changes to the original set of data. This is an undesirable quality. Chap 2

<u>Variable</u> – Any characteristic that can vary. Chap 2

<u>Variability</u> – How much scores of a sample or population differ or deviate from each other. Chap 2

<u>Variable view</u> – SPSS window in which variables are defined. Chap 5

<u>Variance</u> – A measure of variability; the average of the sum of the squared deviations of scores from their mean. The symbol for the population variance is σ^2. Chap 3

<u>Whisker</u> – In a boxplot, a line extending from an edge of the box (either the 25th or 75th percentiles) to the limits of the data. The two whiskers thus extend as far as the range of the data. Chap 3

<u>Y intercept</u> – One of the two determinants of the equation for a straight line. It is the value of Y when X is equal to 0. It is, therefore, the value of Y when the line crosses the Y axis. It has the symbol 'a' in the equation $Y = bX + a$. Chap 14

<u>z score</u> – A conversion of raw data so that the deviation is measured in standard deviation units and the sign, positive or negative, indicates the direction of the deviation. Chap 4

Appendix I – 2
Terms with Different Definitions in Statistics and in Common Usage

Term	Statistical Meaning	Common Usage
Alpha	*Another term for Type I error.*	*The beginning of something.*
Error	*An outcome due to chance, or the variability not due to treatment.*	*A mistake, or the belief in something untrue.*
Level	*With an ANOVA, the number of values of an independent variable.*	*Horizontal, flat, or calm.*
Manipulate	*The researcher determines which condition of the independent variable each subject receives.*	*To exert undue control over someone.*
Power	*The probability of correctly rejecting a false null hypothesis.*	*The ability to act, or having strength or authority.*
Significant	*An outcome is unlikely to have occurred by chance.*	*Full of meaning, important.*
Skewed	*Degree to which one tail in a distribution is larger than the other, and thus the degree to which a distribution is not symmetrical.*	*Slanted or distorted.*
Treatment	*Another term for the independent variable.*	*Medical care, or act of treating.*
Whisker	*In a boxplot, a line extending from an edge of the box to the limits of the data.*	*A hair on a person's face.*

Appendix J
Answers to Chapter/Appendix Problems

Ques								Chap							
	1		2		3		4		5		6		7		8
1	C		D		C		A		B		C		A		B
2	C		C		C		C		C		D		A		C
3	A		D		A		B		A		A		A		E
4	A		B		B		D		D		B		A		A
5	C		A		D		A		A		A		C		B
6	A		A		C		C		B		A		B		A
7	B		B		A		D		C		C		C		E
8	C		B		B		B		A		B		D		C
9	B		A		C		A		B		C		B		A
10	B		B		B		D		C		C		E		B
11	C		C		A		B				D		D		C
12	C		D		D		C				A		A		A
13	A		A		A		D				D		B		A
14	C		A		B		A				B		B		B
15	B		C		C		A				B		C		B
16	A		B		C		C				B		E		C
17	D		C		C		B				A		E		B
18	B		C		C		B				D		D		A
19	B		C		B		A				C		A		A
20	C		A		B		D				A		E		B
21	C				C								B		C
22	D				D								D		A
23	A				B								E		A
24					D								A		
25					B								C		
26					D								A		
27					A								B		
28					C								B		
29					A								C		
30					D								E		
31					A										
32					B										
33					C										
34					B										
35					B										
36					A										
37					B										
38					C										

Ques						Chap						
	9		10		11		12		13		14		15
1	D		C		C		B		C		B		D
2	B		B		C		A		B		C		D
3	B		A		B		D		A		D		B
4	C		B		A		A		C		A		C
5	C		C		B		C		C		B		C
6	B		B		C		D		C		B		A
7	C		A		C		A		A		C		B
8	B		A		B		D		C		A		C
9	A		C		A		C		B		B		D
10	C		B		A		D		A		C		A
11	D		B		B		A		B		B		C
12	C		C		A		B		B		D		
13	A		A		D		B		D		A		
14	B		A		B		A		B		C		
15	B		B		B		D		D		B		
16	A		D		C		C		C		B		
17	B		A		A		C		D		D		
18	A		D		D		A		C		A		
19	A		C		B		D		A		B		
20	B		B		B		D		B		A		
21	D		C		A		D		D		A		
22	C		A		B				C		B		
23	B		B		B				B		C		
24	A		E		C				A		D		
25	B		B		D						A		
26	B		A		B						A		
27	B		B		A						A		
28	A		B		C						C		
29	B		B		A						A		
30	B		C								D		
31	C		B								D		
32	B		D								C		
33	E		B								A		
34	A		C								C		
35	A		B								A		
36	C		C								D		
37	A		B								B		
38	D		C								A		
39	A		A								C		
40			A								B		
41			D										
42			B										

Ques							Appendix				
	‘A’		‘B’		‘C’		‘D’		‘E’		‘L’
1	B		B		B		C		A		A
2	A		C		B		B		E		B
3	D		D		A		B		D		C
4	C		A		D		A		D		A
5	A		C		E		B		C		A
6	C		B		B		E		B		A
7	A		B		A		C		A		B
8	C		C		C		D		C		C
9	D				D		A		A		B
10	B				A		B		B		C
11	C				A						A
12	D										C
13	D										D
14	A										B
15	B										A
16											A
17											B
18											D
19											B
20											C

Appendix K
Statistical Tables

	Table	Chapter/Appendix In which the Table is First Introduced
1	z	Chap 4
2	Chi-square	Chap 7
3	t	Chap 9
4	F	Chap 11
5	q	Chap 11
6	Pearson r	Chap 14
7	Spearman r	Appendix C

Note: For ease of use the entries in the tables have been rounded to two places and limited degrees of freedom are included. If increased accuracy is desired more extensive tables are commonly available. Alternatively, use of a statistical package, such as SPSS, is encouraged.

Table 1a: z Table

Proportion of Curve Below ***Negative*** Values of z

z Scores

z Scores

z	-.00	-.01	-.02	-.03	-.04	-.05	-.06	-.07	-.08	-.09
-2.5	.01	.01	.01	.01	.01	.01	.01	.01	.00	.00
-2.4	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01
-2.3	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01
-2.2	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01
-2.1	.02	.02	.02	.02	.02	.02	.02	.02	.01	.01
-2.0	.02	.02	.02	.02	.02	.02	.02	.02	.02	.02
-1.9	.03	.03	.03	.03	.03	.03	.025	.02	.02	.02
-1.8	.04	.04	.03	.03	.03	.03	.03	.03	.03	.03
-1.7	.04	.04	.04	.04	.04	.04	.04	.04	.04	.04
-1.6	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05
-1.5	.07	.07	.06	.06	.06	.06	.06	.06	.06	.06
-1.4	.08	.08	.08	.08	.07	.07	.07	.07	.07	.07
-1.3	.10	.10	.09	.09	.09	.09	.09	.09	.08	.08
-1.2	.12	.11	.11	.11	.11	.11	.10	.10	.10	.10
-1.1	.14	.13	.13	.13	.13	.13	.12	.12	.12	.12
-1.0	.16	.16	.15	.15	.15	.15	.14	.14	.14	.14
-0.9	.18	.18	.18	.18	.17	.17	.17	.17	.16	.16
-0.8	.21	.21	.21	.20	.20	.20	.20	.19	.19	.19
-0.7	.24	.24	.24	.23	.23	.23	.22	.22	.22	.22
-0.6	.27	.27	.27	.26	.26	.26	.26	.25	.25	.25
-0.5	.31	.31	.30	.30	**.30**	.29	.29	.28	.28	.28
-0.4	.34	.34	.34	.33	.33	.33	.32	.32	.32	.31
-0.3	.38	.38	.38	.37	.37	.36	.36	.36	.35	.35
-0.2	.42	.42	.41	.41	.41	.40	.40	.39	.39	.39
-0.1	.46	.46	.45	.45	.44	.44	.44	.43	.43	.43
-0.0	.50	.50	.49	.49	.48	.48	.48	.47	.47	.46

Proportions

Example: Proportion of the curve below a z of –0.54 equals .30, which is bolded and underlined in the table.

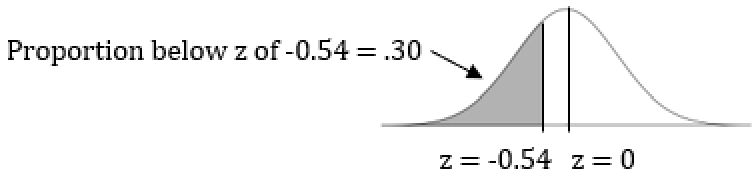


Table 1b: z Table

Proportion of Curve Below *Positive* Values of z

z Scores

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.50	.50	.51	.51	.52	.52	.52	.53	.53	.54
0.1	.54	.54	.55	.55	.56	.56	.56	.57	.57	.58
0.2	.58	.58	.59	.59	.59	.60	.60	.61	.61	.61
0.3	.62	.62	.63	.63	.63	.64	.64	.64	.65	.65
0.4	.66	.66	.66	.67	.67	.67	.68	.68	.68	.69
0.5	.69	.70	.70	.70	.71	.71	.71	.72	.72	.72
0.6	.73	.73	.73	.74	.74	.74	.75	.75	.75	.76
0.7	.76	.76	.76	.77	.77	.77	.78	.78	.78	.79
0.8	.79	.79	.79	.80	.80	.80	.80	.81	.81	.81
0.9	.82	.82	.82	.82	.83	.83	.83	.83	.84	.84
1.0	.84	.84	.85	.85	.85	.85	.86	.86	.86	.86
1.1	.86	.87	.87	.87	**.87**	.88	.88	.88	.88	.88
1.2	.89	.89	.89	.89	.89	.89	.90	.90	.90	.90
1.3	.90	.91	.91	.91	.91	.91	.91	.92	.92	.92
1.4	.92	.92	.92	.92	.93	.93	.93	.93	.93	.93
1.5	.93	.94	.94	.94	.94	.94	.94	.94	.94	.94
1.6	.95	.95	.95	.95	.95	.95	.95	.95	.95	.96
1.7	.96	.96	.96	.96	.96	.96	.96	.96	.96	.96
1.8	.96	.97	.97	.97	.97	.97	.97	.97	.97	.97
1.9	.97	.97	.97	.97	.97	.97	.975	.98	.98	.98
2.0	.98	.98	.98	.98	.98	.98	.98	.98	.98	.98
2.1	.98	.98	.98	.98	.98	.98	.99	.99	.99	.99
2.2	.99	.99	.99	.99	.99	.99	.99	.99	.99	.99
2.3	.99	.99	.99	.99	.99	.99	.99	.99	.99	.99
2.4	.99	.99	.99	.99	.99	.99	.99	.99	.99	.99
2.5	.99	.99	.99	.99	1.0	1.0	1.0	1.0	1.0	1.0

z Scores

Proportions

Example: Proportion of the curve below a z of 1.14 equals .87, which is bolded and underlined in the table.

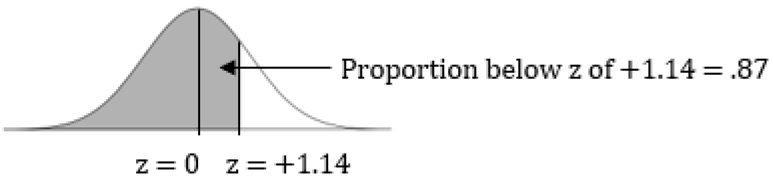


Table 2: Critical Values for the Chi-Square Test*

df	α = .05	α = .01
1	3.84	6.64
2	5.99	9.21
3	**<u>7.82</u>**	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.72
12	21.03	26.22
13	22.36	27.69
14	23.68	29.14
15	25.00	30.58
16	26.30	32.00
17	27.59	33.41
18	28.87	34.80
19	30.14	36.19
20	31.41	37.57

***To reject the null hypothesis, your calculated value for the Chi-Square test must be larger than the critical value in the table.**

Example: With α = .05 and 3 df the critical value is 7.82, which is bolded and underlined in the table.

Table 3a: Critical Values for the *One*-tailed t Test*

df	α = .05	α = .01
1	6.31	31.82
2	2.92	6.97
3	2.35	4.54
4	2.13	**3.75**
5	2.02	3.37
6	1.94	3.14
7	1.90	3.00
8	1.86	2.90
9	1.83	2.82
10	1.81	2.76
11	1.80	2.72
12	1.78	2.68
13	1.77	2.65
14	1.76	2.63
15	1.75	2.60
16	1.75	2.58
17	1.74	2.57
18	1.73	2.55
19	1.73	2.54
20	1.73	2.53
30	1.70	2.46
60	1.67	2.39
100	1.66	2.36
∞	1.65	2.33

***To reject the null hypothesis, the absolute value of your calculated t test must be larger than the critical value in the table, and the outcome must be in the predicted direction.**

Example: With α = .01 and 4 df the critical value is 3.75, which is bolded and underlined in the table. As this is a one-tailed test the null hypothesis will determine whether the critical value is +3.75 or −3.75.

Table 3b: Critical Values for the *<u>Two</u>*-tailed t Test*

df	α = .05	α = .01
1	12.71	63.66
2	4.30	9.93
3	3.18	5.84
4	2.78	**<u>4.60</u>**
5	2.57	4.03
6	2.45	3.71
7	2.37	3.50
8	2.31	3.36
9	2.26	3.25
10	2.23	3.17
11	2.20	3.11
12	2.18	3.06
13	2.16	3.01
14	2.15	2.98
15	2.13	2.95
16	2.12	2.92
17	2.11	2.90
18	2.10	2.88
19	2.09	2.86
20	2.08	2.85
30	2.04	2.75
60	2.00	2.66
100	1.98	2.63
∞	1.96	2.58

***To reject the null hypothesis, the absolute value of your calculated t test must be larger than the critical value in the table.**

Example: With α = .01 and 4 df the critical value is 4.60, which is bolded and underlined in the table. As this is a two-tailed test, in order to be statistically significant your calculated value for the t test must be less than **–**4.60 or greater than +4.60.

Table 4: Critical Values for the F Test*

Alpha = .05

	df numerator					
df denominator ↓	**1**	**2**	**3**	**4**	**5**	**6**
1	161.45	199.50	215.71	224.58	230.16	233.99
2	18.51	19.00	19.16	19.25	19.30	19.33
3	10.13	9.55	9.28	9.12	9.01	8.94
4	7.71	6.94	6.59	6.39	6.26	6.16
5	6.61	5.79	5.41	5.19	5.05	4.95
6	5.99	5.14	4.76	4.53	4.39	4.28
7	5.59	4.74	4.35	4.12	3.97	3.87
8	5.32	4.46	4.07	3.84	3.69	3.58
9	5.12	4.26	3.86	3.63	3.48	3.37
10	4.96	4.10	3.71	3.48	3.33	3.22
11	4.84	3.98	3.59	3.36	3.20	3.09
12	4.75	3.89	3.49	3.26	3.11	3.00
13	4.67	3.81	3.41	3.18	3.03	2.92
14	4.60	3.74	3.34	3.11	2.96	2.85
15	4.54	3.68	3.29	**3.06**	2.90	2.79
16	4.49	3.63	3.24	3.01	2.85	2.74
17	4.45	3.59	3.20	2.96	2.81	2.70
18	4.41	3.55	3.16	2.93	2.77	2.66
19	4.38	3.52	3.13	2.90	2.74	2.63
20	4.35	3.49	3.10	2.87	2.71	2.60
21	4.32	3.47	3.07	2.84	2.68	2.57
22	4.30	3.44	3.05	2.82	2.66	2.55
23	4.28	3.42	3.03	2.80	2.64	2.53
24	4.26	3.40	3.01	2.78	2.62	2.51
25	4.24	3.38	2.99	2.76	2.60	2.49
30	4.17	3.32	2.92	2.69	2.53	2.42
40	4.08	3.23	2.84	2.61	2.45	2.34
60	4.00	3.15	2.76	2.52	2.37	2.25
80	3.96	3.11	2.72	2.48	2.33	2.21
100	3.94	3.09	2.70	2.46	2.30	2.19
150	3.91	3.06	2.67	2.43	2.27	2.16

***To reject the null hypothesis, the value of your calculated F ratio must be larger than the critical value in the table.**

Example: With 4 df in the numerator of the F ratio and 15 df in the denominator, the critical value equals 3.06. This value is bolded and underlined in the table.

Table 5: Values of q for Tukey HSD Test used with Main Effects*

Alpha = .05

Number of Levels of the Independent Variable (k).

Alternatively, the total number of means being compared.

df for denominator of F ratio ↓	3	4	5	6
5	4.60	5.22	5.67	6.03
6	4.34	4.90	5.30	5.63
7	4.16	4.68	5.06	5.36
8	4.04	4.53	4.89	5.17
9	3.95	4.41	4.76	5.02
10	3.88	**<u>4.33</u>**	4.65	4.91
11	3.82	4.26	4.57	4.82
12	3.77	4.20	4.51	4.75
13	3.73	4.15	4.45	4.69
14	3.70	4.11	4.41	4.64
15	3.67	4.08	4.37	4.59
16	3.65	4.05	4.33	4.56
17	3.63	4.02	4.30	4.52
18	3.61	4.00	4.28	4.49
19	3.59	3.98	4.25	4.47
20	3.58	3.96	4.23	4.45
21	3.57	3.94	4.21	4.43
22	3.56	3.93	4.20	4.41
23	3.54	3.92	4.18	4.39
24	3.53	3.90	4.17	4.37
25	3.52	3.89	4.16	4.36
30	3.49	3.85	4.10	4.30
40	3.44	3.79	4.04	4.23

***The entries in this Table are NOT the critical values for the Tukey HSD test. The entries in this Table are used to calculate the critical value of the Tukey HSD test.**

Example: With a one-way between-subjects ANOVA with 4 levels for the IV, there are a total of 4 sample means being compared and if there are 10 df in the denominator of the F ratio, q equals 4.33. This value is bolded and underlined in the table.

Table 6: Critical Values for the *Two*-tailed Pearson Correlation*

df**	α = .05	α = .01
1	1.00	1.00
2	.95	.99
3	.88	.96
4	.81	.92
5	.75	.87
6	**.71**	.83
7	.67	.80
8	.63	.77
9	.60	.74
10	.58	.71
11	.55	.68
12	.53	.66
13	.51	.64
14	.50	.62
15	.48	.61
16	.47	.59
17	.46	.58
18	.44	.56
19	.43	.55
20	.42	.54
25	.37	.49
30	.35	.45
50	.27	.35
70	.23	.30
100	.20	.25

***To reject the null hypothesis, the absolute value of your calculated Pearson r must be larger than the critical value in the table.**

**df = n – 2 where n equals the number of pairs of scores

Example: With α = .05 and 6 df, the critical value equals .71. This value is bolded and underlined in the table. As this is a two-tailed test, in order to be statistically significant the calculated value of your Pearson r must be less than −.71 or greater than +.71.

Table 7: Critical Values for the *Two*-tailed Spearman Correlation*

Number of Pairs of Scores	**α = .05**	**α = .01**
5	1.00	----
6	.89	1.00
7	.79	.93
8	.74	.88
9	.70	.83
10	.65	.79
11	.62	.76
12	**.59**	.73
13	.56	.70
14	.54	.68
15	.52	.65
16	.50	.64
17	.49	.62
18	.47	.60
19	.46	.58
20	.45	.57
25	.40	.51
30	.36	.47
50	.28	.36
70	.24	.31
100	.20	.26

***To reject the null hypothesis, the absolute value of your calculated Spearman r must be as large, or larger, than the critical value in the table.**

Example: With α = .05 and with 12 pairs of scores the critical value is .59. This value is bolded and underlined in the table. As this is a two-tailed test, in order to be statistically significant the calculated value of your Spearman r must be less than or equal to −.59 or greater than or equal to +.59.

Appendix L
Overview of Statistical Procedures Covered in This Book

Descriptive Procedures (Summarizing Sample Data)

	Type of Data			
	Nominal (Frequency)	Ordinal (Ranked)	Interval/Ratio (Score)	
Frequency Dist	Bar Graph or Pie Chart	Bar Graph	Histogram or Frequency Polygon	
Central Tendency	Mode	Median	IF NOT NORMAL Median	IF NORMAL Mean (Median – less common)
Variability	– – – –	Range	Interquartile Range	Standard Deviation z Score
Summary Presentation			Stem-and-leaf display and Boxplot	Stem-and-leaf display and Boxplot

Dist, distribution

Inferential Procedures (Focus is on Statistical Significance: Making Inferences About Population Characteristics)

Type of Data

Nominal (Frequency)	Ordinal (Ranked)	Interval/Ratio (Continuous Measure)

When the Focus is on the Statistical Significance of a Difference:

Research Design		Research Design		
One Variable With At Least Two Outcomes	Goodness-of-fit Chi-Square	One IV With One Sample		One-sample z Test or One-sample t Test
		One IV With Two Or More Independent Samples	Kruskal–Wallis H	One-way Between–Subjects ANOVA (Only two independent samples, Independent Samples t Test)
		One IV With One Sample Having Two Or More Repeated Measures		One-way Within–Subjects ANOVA (Only two repeated measures, Dependent Samples t Test)
Two Variables, Each With At Least Two Outcomes	Chi-Square Test of Independence	Two IV Each With Two Or More Independent Samples		Two-way Between–Subjects ANOVA

When the focus is on Characteristics and Statistical Significance of an Association:

Research Question			
Association:	Chi-Square Test of Independence		
Correlation:	Phi r	Spearman r	Pearson r Multiple Correlation
Regression:			Regression Multiple Regression

IV, independent variable; ANOVA, analysis of variance.

Appendix M
Comparison of ANOVAs

ONE-WAY WITHIN-SUBJECTS		***ONE-WAY BETWEEN- SUBJECTS***		*TWO-WAY BETWEEN-SUBJECTS*	
Source	F ratio	**Source**	**F ratio**	Source	F ratio
Between Treatments	MS_{Bet}/MS_{Res}	**Between Groups**	**MS_{Bet}/MS_{W}**	Partitioned into:	
				Factor A	MS_{A}/MS_{W}
				Factor B	MS_{B}/MS_{W}
				Interaction AXB	MS_{AXB}/MS_{W}
Within Treatments is Partitioned into: ~~Pre-existing subject diff~~ and Residual error		**Within Groups**		Within Groups	
Total		**Total**		Total	

If No F ratio Is Statistically Significant Your Analysis Is Complete

If At Least One F Ratio Is Statistically Significant

ONE-WAY WITHIN-SUBJECTS		***ONE-WAY BETWEEN-SUBJECTS***		*TWO-WAY BETWEEN-SUBJECTS*	
WHERE IS EFFECT?	*HOW BIG IS EFFECT?*	***WHERE IS EFFECT?***	***HOW BIG IS EFFECT?***	*WHERE IS EFFECT?*	*HOW BIG IS EFFECT?*
Post hoc test: Dependent t tests using Bonferroni method	Partial eta squared (η_p^2)	**Post hoc test: Tukey HSD** **Need to find q**	**Eta squared (η^2)**	Post hoc test: Tukey HSD	Eta squared (η^2), Partial eta squared (η_p^2), or both
				If sig main effect for A, need to find q	If sig main effect for A, report η^2 or η_p^2 or both
				If sig main effect for B, need to find q	If sig main effect for B, report η^2 or η_p^2 or both
				If sig interaction AXB, need to use q_i	If sig interaction AXB, report η^2 or η_p^2 or both

Appendix N
Choosing the Correct Inferential Procedure

Looking for a Difference or Interaction

		Statistical Procedure	*Effect Size*	*Post Hoc*
Nominal Data	One Variable	Goodness of Fit χ^2	—	—
	Two Variables	χ^2 Test of Independence	Phi / Cramer's V	χ^2 with Bonferroni
Interval/Ratio Data	One IV with one Sample	One sample z test (or confidence interval)	—	—
		One sample t test (or confidence interval)	eta^2	—
	One IV with 2 independent samples	Independent samples t test	eta^2	—
	One IV with 2 or more independent samples	One-way between-subjects ANOVA	eta^2	Tukey HSD
	One IV with 1 sample and 2 repeated measures (or matched samples)	Dependent samples t test	eta^2	—
	One IV with 1 sample and 2 or more repeated measures (or matched samples)	One-way within-subjects ANOVA	eta^2 or partial eta^2	Dependent t tests with Bonferroni
	Two IV each with 2 or more independent samples	Two-way between-subjects ANOVA	eta^2 or partial eta^2 or both	Tukey HSD

Looking for an Association or Correlation

		Statistical Procedure	*Effect Size*	*Then*
Interval/Ratio Data	Two variables	Pearson r	r^2	Linear Regression

Practice Choosing the Correct Procedure

(Answers are provided in Appendix J)

1. We wish to determine whether a die is fair.
   a. Goodness-of-fit chi-square
   b. Independent t test or one-way between-subjects ANOVA
   c. Pearson r
   d. Two-way between-subjects ANOVA

2. Each subject is randomly assigned to one of five levels of studying and then their exam grades are compared.
   a. Goodness-of-fit chi-square
   b. One-way between-subjects ANOVA
   c. Pearson r
   d. Two-way between-subjects ANOVA

3. We measure how tall each person is and then look to see if there is an association with how high they can jump.
   a. Two-way between-subjects ANOVA
   b. One-way between-subjects ANOVA
   c. Pearson r
   d. One-sample z

4. We compare the GPA's (grade point averages) of men versus women who have, or have not, studied abroad.
   a. Two-way between-subjects ANOVA
   b. Independent t test or one-way between-subjects ANOVA
   c. Pearson r
   d. Goodness-of-fit chi-square

5. We check the claim that a person can flip a coin so it tends to land heads.
   a. Goodness-of-fit chi-square
   b. One-way between-subjects ANOVA
   c. Pearson r
   d. One-sample z

6. We compare the frequencies of social science majors and humanities majors, and their choice of political party preference (democratic, republican, other).
   a. Chi-square test of independence
   b. One-way between-subjects ANOVA
   c. Pearson r
   d. Two-way between-subjects ANOVA

7. We examine if there is a difference between whether a student is married or not and how happy they report they are on a 25-item scale.
   a. Pearson r
   b. Independent t test or one-way between-subjects ANOVA
   c. Goodness-of-fit chi-square
   d. Two-way between-subjects ANOVA

8. A researcher is interested in whether there is a difference between a person's gender and whether they vote democratic or republican.
   a. Pearson r

b. One-way between-subjects ANOVA
c. Chi-square test of independence
d. One-way within-subjects ANOVA

9. A researcher examines whether living at high altitudes affects IQ. She compares the IQ data from 500 people who live at high altitudes to the known mean and standard deviation for the general population.
   a. One-way between-subjects ANOVA
   b. One-sample z
   c. One-sample t
   d. Chi-square test of independence

10. We compare the age of death for people who had a pet versus those who did not have a pet.
   a. Pearson r
   b. Two-way between-subjects ANOVA
   c. Independent samples t test or one-way between-subjects ANOVA
   d. One-sample t

11. A study for a car magazine examines whether there is an association between the weight of a car and how many feet it takes for it to stop from 60 mph. Data are collected from 30 cars of various weights.
   a. Pearson r
   b. One-way between-subjects ANOVA
   c. Chi-square test of independence
   d. One-sample t

12. From past history it is known that with a particular manufacturing process 10% of the product has been defective. A new process is instituted and for the first 100 items there are only 6 that are defective. Has the frequency of defective product been reduced?
   a. One-sample t
   b. Pearson r
   c. Goodness-of-fit chi-square
   d. One-way between-subjects ANOVA

13. A teacher is interested in whether there is an association between gender (male or female) and openness to experience (measured on a 25-point scale).
   a. One-sample z
   b. Pearson r
   c. One-sample t
   d. One-way between-subjects ANOVA

14. A restaurant wants to determine whether the quality of their five most popular offerings differ according to reviewers. Ten food tasters are invited to give each dish a rating from 1 to 10.
   a. One-way between-subjects ANOVA
   b. One-way within-subjects ANOVA
   c. Two-way between-subjects ANOVA
   d. Pearson r

15. A physical education instructor compares males and females and whether they prefer playing basketball or volleyball.
   a. Chi-square test of independence
   b. One-way within-subjects ANOVA
   c. One-way between-subjects ANOVA

d. Pearson r

16. We compare males and females who are judged to be either attractive or not on how outgoing they are (measured on a 15–point scale).
    a. One-way between-subjects ANOVA
    b. One-way within-subjects ANOVA
    c. Two-way between-subjects ANOVA
    d. One-sample t

17. A newspaper examines whether there is an association between age and the number of speeding tickets received by 100 drivers over the previous 3 years.
    a. Independent samples t test or one-way between-subjects ANOVA
    b. Pearson r
    c. Chi-square test of independence
    d. Dependent samples t test or one-way within subjects ANOVA

18. A researcher checks to see if there is a difference between handedness (either left or right) and grade point average.
    a. Chi-square test of independence
    b. Pearson r
    c. One-sample t
    e. Independent samples t test or one-way between-subjects ANOVA

19. A study is conducted which examines whether there is a difference in the type of car (domestic or foreign) driven by republicans and democrats.
    a. One-sample t
    b. Chi-square test of independence
    c. One-way between-subjects ANOVA
    d. One-way within-subjects ANOVA

20. At a college a study is conducted that compares whether appreciation of the liberal arts (measured on a 20-item scale) is affected by major (art or science) and class in college (freshman, sophomore, junior, or senior).
    a. One-way between-subjects ANOVA
    b. One-way within-subjects ANOVA
    c. Two-way between-subjects ANOVA
    d. Pearson r

References

Bandura, A., Ross, D., & Ross, S. A. (1961). Transmission of aggression through imitation of aggressive models. *Journal of Abnormal and Social Psychology, 63,* 575-582. (Chap 6)

Boorstin, D. J. (1983). *The discoverers.* New York: Random House. (Chap 12, 20, 23, 30, 32)

Brown, M. B. & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association, 69,* 364-367. (Chap 11, in SPSS)

Caspi, A., McClay, J., Moffitt, T., Mill, J., Martin, J., Craig, I. W., Taylor, A., Poulton, R., & Moffitt, T. (2002). Role of genotype in the cycle of violence in maltreated children. *Science, 297,* 851 – 854. (Chap 13)

Chou, K. L., Ho, A. H. Y., & Chi, I. (2006). Living alone and depression in Chinese older adults. *Aging & Mental Health*, *10*(6), 583-591. (Chap 8)

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. 2nd Ed., Lawrence Erlbaum Associates, Hillsdale, NJ. (Chap 8 and App IV)

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155 –159. (App IV)

Feinberg, W. E. (1971). Teaching type I and type II errors: The judicial process. *The American Statistician, 25*(3), 30-32. (Chap 6)

Field, A. (2009). *Discovering statistics using SPSS, 3rd edition.* Los Angeles: SAGE. (Chap 9)

Fisher, R. A. (1971) (1935). The design of experiments (9th ed.). Macmilan. (Chap 15)

Gazzaniga, M. S. (1967). The split brain in man. *Scientific American, 217* (2), 24 -29. (Chap 6)

Gould, S. J. (1981). *The mismeasure of man.* New York: Norton. (Chap 16)

Levine, T. R. & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research, 28,* 612-625. (Chap 13)

Lumley, T., Diehr, P., Emerson, S. & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health, 23*, 151-169. (Chap 9)

Martin, M. A. (2003). "It's like … you know": The use of analogies and heuristics in teaching introductory statistical methods. *The Journal of Statistics Education, 11*(2) (www.amstat.org/publications/jse/v11n2/martin.html) (Chap 6)

Mathes, E. (2003). Are sex differences in sexual vs emotional jealousy explained better by differences in sexual strategies or uncertainty of paternity. *Psychological Reports, 93*(3), 895 –906. (Chap 8)

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*(2), 147–163. (App IV)

Norvilitis, J. M. & Reid, H. M. (2011, March). *College success: The relations between* *appreciation of the liberal arts, symptoms of ADHD and study skills*. Poster presented at the Eastern Psychological Association Convention, Cambridge, MA. (App III)

Perfect, T. (2003). Local processing bias impairs lineup performance. *Psychological Reports, 93*(2), 393 –394. (Chap 8)

Sandson, T. A., Bachna, K. J. & Morin, M. D. (2000). Right hemisphere dysfunction in ADHD: Visual hemispatial inattention and clinical subtype. *Journal of Learning Disabilities, 33*(1), 83 - 90. (Chap 8)

Index