Statistics Minus The Math

An Introduction for the Social Sciences

Nathan Favero

2024-07-18

Get the latest version of this text at: https://minusthemath.com

Table of contents

Preface

This book was largely adapted from the public domain resource *Online Statistics Education: A Multimedia Course of Study* (http://onlinestatbook.com/ Project Leader: David M. Lane, Rice University). A huge thanks to David Lane and his colleagues at Rice University for their creation of this wonderful resource. I use footnotes throughout to indicate precisely where the various sections of each chapter came from. Chapters 11-13 (as well as Section 4.4 and Section 3.5) were written by me (Nathan).

This book is meant to be a free resource and is licensed under CC BY 4.0. You're welcome to share or adapt it, so long as you provide attribution to any work of mine that you use. This book was made using Quarto and is hosted using GitHub Pages.

There are still formatting inconsistencies, and this is ever a work in progress. If you find errors, feel free to reach out (find updated contact info here: https://nathanfavero.com) so I can correct them for the next version I publish.

What's Unique About This Text?

1. It is a true introduction, not assuming any prior training in statistics.
2. I try to minimize use of math (beyond the very elementary), instead focusing on conceptual description and interpretation.
3. I introduce regression very early on (Ch. 3) so that students (especially PhD students) can quickly get started on their term papers and better understand any quantitative articles they're reading. The treatment of regression is further built out in the final chapters (11-13). Regression is, after all, the workhorse of applied statistics for the social sciences.
4. I skip a traditional treatment of probability theory because I don't find traditional treatments to be very useful for students interested in applied statistics. Instead, I've written a chapter (11) on the logic and practice of building models that account for uncertainty.
5. There is a bit of Stata code in the final chapter, but otherwise all examples are presented apart from any statistical software package.

Past Versions

This version (1.3) was updated July 18, 2024. PDFs of past versions are available at https://minusthemath.com.

- Version 1.3 updates: Added material on multiple regression (end of Chapter 3) and a new Section 4.4 on interpreting confidence intervals. Various formatting updates. Notation is updated in line with conventions: regression parameters are redone, and $\bar{X}$ is now used for sample mean and n for sample size.
- Version 1.2 updates: The discussion of transforming variables now appears in Chapter 2 (rather than Chapter 3).

1 Graphical Tools for Describing Variables

1.1 Variables[1]

Variables are properties or characteristics of some event, object, or person that can take on different values or amounts (as opposed to constants such as π that do not vary). When conducting research, experimenters often manipulate or measure variables. For example, an experimenter might compare the effectiveness of four types of antidepressants. In this case, the variable is "type of antidepressant." This experimenter might also ask study participants to indicate their mood on a scale of 1 to 10. "Mood" would be a second variable.

1.1.1 Qualitative and Quantitative Variables

An important distinction between variables is between qualitative variables and quantitative variables. **Qualitative variables** are those that express a qualitative attribute such as hair color, eye color, religion, favorite movie, gender, and so on. The values of a qualitative variable do not imply a numerical ordering. Values of the variable "religion" differ qualitatively; no ordering of religions is implied. Qualitative variables are also sometimes referred to as categorical or nominal variables. **Quantitative variables** are those variables that are measured in terms of numbers. Some examples of quantitative variables are height, weight, and shoe size.

> Example: Can blueberries slow down aging? A study indicates that antioxidants found in blueberries may slow down the process of aging. In this study, 19-month-old rats (equivalent to 60-year-old humans) were fed either their standard diet or a diet supplemented by either blueberry, strawberry, or spinach powder. After eight weeks, the rats were given memory and motor skills tests. Although all supplemented rats showed improvement, those supplemented with blueberry powder showed the most notable improvement.
>
> More information: https://www.apa.org/monitor/dec01/blueberries.html

[1]This section is adapted from Heidi Ziemer. "Variables." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/introduction/variables.html

In the study on the effect of diet discussed above, the independent variable was type of supplement: none, strawberry, blueberry, and spinach. The variable "type of supplement" is a qualitative variable; there is nothing quantitative about it. In contrast, the dependent variable "memory test" is a quantitative variable since memory performance was measured on a quantitative scale (number correct).

1.1.2 Discrete and Continuous Variables

Variables such as number of children in a household are called **discrete variables** since the possible scores are discrete points on the scale. For example, a household could have three children or six children, but not 4.53 children. Other variables such as "time to respond to a question" are **continuous variables** since the scale is continuous and not made up of discrete steps. The response time could be 1.64 seconds, or it could be 1.64237123922121 seconds. Of course, the practicalities of measurement preclude most measured variables from being truly continuous.

1.2 Percentiles[2]

Before turning our attention to some basic graphical tools we use to visualize qualitative and quantitative variables, it is helpful to also briefly go over **percentiles** since they will be used in some of these tools. Many of us have probably encountered percentiles before in the context of standardized exam testing. A test score in and of itself is usually difficult to interpret. For example, if you learned that your score on a measure of shyness was 35 out of a possible 50, you would have little idea how shy you are compared to other people. More relevant is the percentage of people with lower shyness scores than yours. This percentage is called a percentile. If 65% of the scores were below yours, then your score would be the 65th percentile.

1.2.1 Three Alternative Definitions of Percentile

There is no universally accepted definition of a percentile. Using the 65th percentile as an example, the 65th percentile can be defined as the lowest score that is greater than 65% of the scores. This is the way we defined it above and we will call this "Definition 1." The 65th percentile can also be defined as the smallest score that is greater than *or equal to* 65% of the scores. This we will call "Definition 2." Though these two definitions appear very similar, they can sometimes lead to dramatically different results, especially when there is relatively little data. Moreover, neither of these definitions is explicit about how to handle rounding. For instance, what rank is required to be higher than 65% of the scores when the total number of

[2]This section is adapted from David M. Lane. "Percentiles." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/introduction/percentiles.html

scores is 50? This is tricky because 65% of 50 is 32.5. How do we find the lowest number that is higher than 32.5 of the scores?

A third way to compute percentiles is a weighted average of the percentiles computed according to the first two definitions. The details of computing percentiles under this third definition are a bit complicated, but fortunately, statistical software can easily do the calculations for us. Since it is unlikely you will need to compute percentiles by hand, we leave the details of these computations to the appendix appearing at the end of this chapter. Despite its complexity, the third definition handles rounding more gracefully than the other two and has the advantage that it allows the median to be defined conveniently as the 50th percentile. Unless otherwise specified, when we refer to "percentile," we will be referring to this third definition of percentiles.

1.3 Graphing Qualitative Variables[3]

When Apple Computer introduced the iMac computer in August 1998, the company wanted to learn whether the iMac was expanding Apple's market share. Was the iMac just attracting previous Macintosh owners? Or was it purchased by newcomers to the computer market and by previous Windows users who were switching over? To find out, 500 iMac customers were interviewed. Each customer was categorized as a previous Macintosh owner, a previous Windows owner, or a new computer purchaser.

This section examines graphical methods for displaying the results of the interviews. We'll learn some general lessons about how to graph data that fall into a small number of categories. A later section will consider how to graph numerical data in which each observation is represented by a number in some range. The key point about the qualitative data that occupy us in the present section is that they do not come with a pre-established ordering (the way numbers are ordered). For example, there is no natural sense in which the category of previous Windows users comes before or after the category of previous Macintosh users. This situation may be contrasted with quantitative data, such as a person's weight. People of one weight are naturally ordered with respect to people of a different weight.

1.3.1 Frequency Tables

All of the graphical methods shown in this section are derived from frequency tables. Table 1.1 shows a frequency table for the results of the iMac study; it shows the frequencies of the various response categories. It also shows the relative frequencies, which are the proportion of responses in each category. For example, the relative frequency for "none" is $85/500 = 0.17$.

[3]This section is adapted from David M. Lane. "Graphing Qualitative Variables." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/graphing_distributions/graphing_qualitative.html

Table 1.1: Frequency Table for the iMac Data.

Previous Ownership	Frequency	Relative Frequency
None	85	0.17
Windows	60	0.12
Macintosh	355	0.71
Total	500	1.00

1.3.2 Pie Charts

The pie chart in Figure 1.1 shows the results of the iMac study. In a pie chart, each category is represented by a slice of the pie. The area of the slice is proportional to the percentage of responses in the category. This is simply the relative frequency multiplied by 100. Although most iMac purchasers were Macintosh owners, Apple was encouraged by the 12% of purchasers who were former Windows users, and by the 17% of purchasers who were buying a computer for the first time.

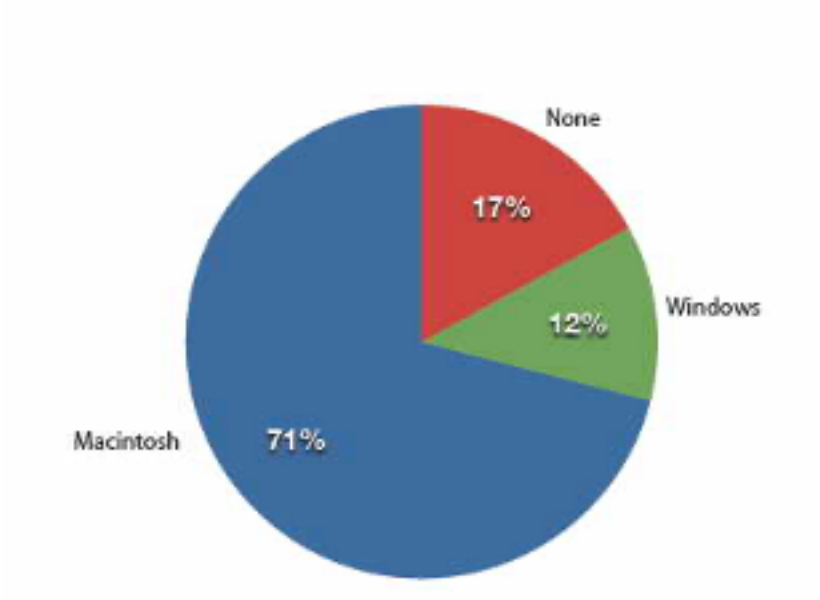


Figure 1.1: Pie chart of iMac purchases illustrating frequencies of previous computer ownership.

Pie charts are effective for displaying the relative frequencies of a small number of categories. They are not recommended, however, when you have a large number of categories. Pie charts can also be confusing when they are used to compare the outcomes of two different surveys or

experiments. In an influential book on the use of graphs, Edward Tufte asserted, "The only worse design than a pie chart is several of them."

Here is another important point about pie charts. If they are based on a small number of observations, it can be misleading to label the pie slices with percentages. For example, if just 5 people had been interviewed by Apple Computers, and 3 were former Windows users, it would be misleading to display a pie chart with the Windows slice showing 60%. With so few people interviewed, such a large percentage of Windows users might easily have occurred since chance can cause large errors with small samples. In this case, it is better to alert the user of the pie chart to the actual numbers involved. The slices should therefore be labeled with the actual frequencies observed (e.g., 3) instead of with percentages.

1.3.3 Bar Charts

Bar charts can also be used to represent frequencies of different categories. A bar chart of the iMac purchases is shown in Figure 1.2. Frequencies are shown on the Y-axis and the type of computer previously owned is shown on the X-axis. Typically, the Y-axis shows the number of observations in each category rather than the percentage of observations as is typical in pie charts.

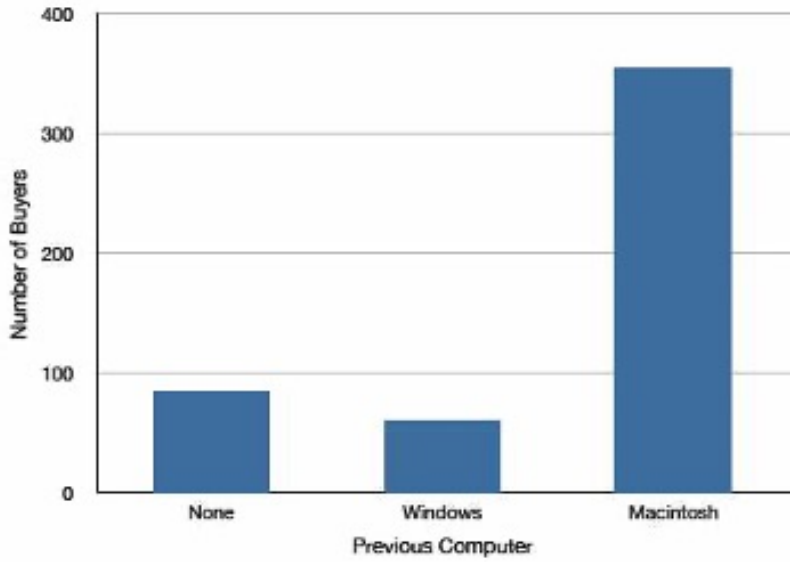


Figure 1.2: Bar chart of iMac purchases as a function of previous computer ownership.

1.3.3.1 Comparing Distributions

Often we need to compare the results of different surveys, or of different conditions within the same overall survey. In this case, we are comparing the "distributions" of responses between the surveys or conditions. Bar charts are often excellent for illustrating differences between two distributions. Figure 1.3 shows the number of people playing card games at the Yahoo website on a Sunday and on a Wednesday in the Spring of 2001. We see that there were more players overall on Wednesday compared to Sunday. The number of people playing Pinochle was nonetheless the same on these two days. In contrast, there were about twice as many people playing hearts on Wednesday as on Sunday. Facts like these emerge clearly from a well-designed bar chart.

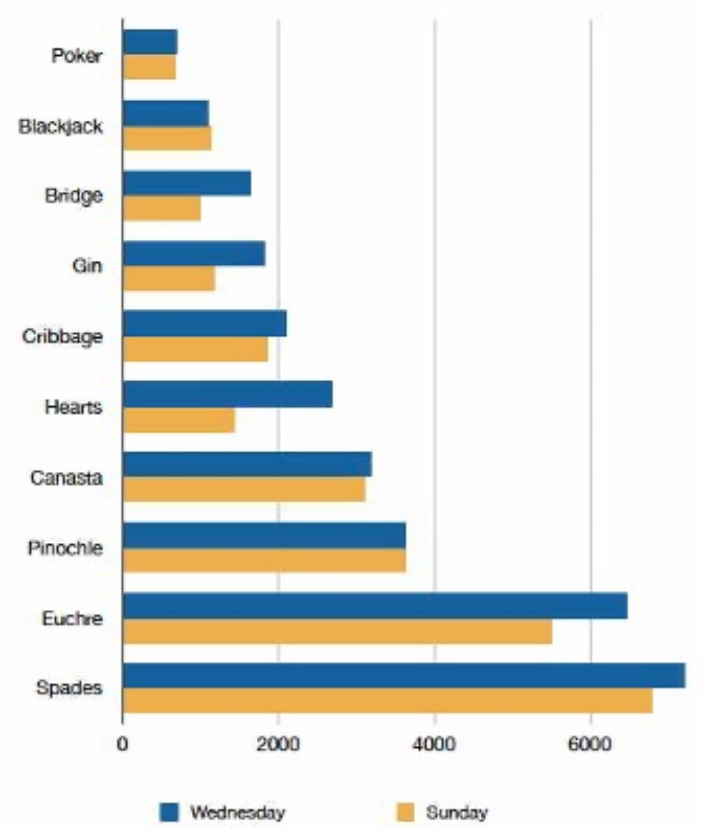


Figure 1.3: A bar chart of the number of people playing different card games on Sunday and Wednesday.

The bars in Figure 1.3 are oriented horizontally rather than vertically. The horizontal format is useful when you have many categories because there is more room for the category labels. We'll have more to say about bar charts when we consider numerical quantities later in the section Section 1.4.4.

1.3.4 Some graphical mistakes to avoid

Don't get fancy! People sometimes add features to graphs that don't help to convey their information. For example, 3-dimensional bar charts such as the one shown in Figure 1.4 are usually not as effective as their two-dimensional counterparts.

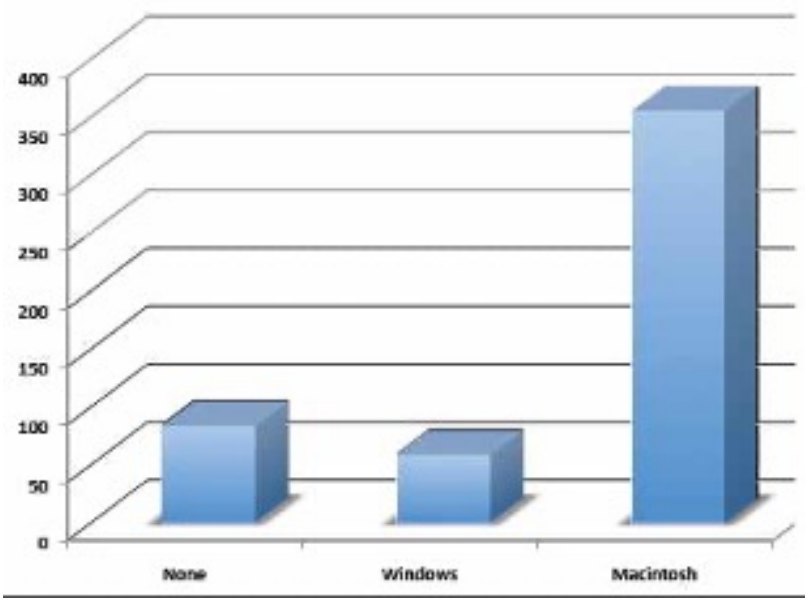


Figure 1.4: A three-dimensional version of Figure 1.2.

Here is another way that fanciness can lead to trouble. Instead of plain bars, it is tempting to substitute meaningful images. For example, Figure 1.5 presents the iMac data using pictures of computers. The heights of the pictures accurately represent the number of buyers, yet Figure 1.5 is misleading because the viewer's attention will be captured by areas. The areas can exaggerate the size differences between the groups. In terms of percentages, the ratio of previous Macintosh owners to previous Windows owners is about 6 to 1. But the ratio of the two areas in Figure 1.5 is about 35 to 1. A biased person wishing to hide the fact that many Windows owners purchased iMacs would be tempted to use Figure 1.5 instead of Figure 1.2! Edward Tufte coined the term "lie factor" to refer to the ratio of the size of the effect shown in a graph to the size of the effect shown in the data. He suggests that lie factors greater than 1.05 or less than 0.95 produce unacceptable distortion.

Another distortion in bar charts results from setting the baseline to a value other than zero. The baseline is the bottom of the Y-axis, representing the least number of cases that could have occurred in a category. Normally, but not always, this number should be zero. Figure 1.6 shows the iMac data with a baseline of 50. Once again, the differences in areas suggest a different story than the true differences in percentages. The percentage of Windows-switchers seems minuscule compared to its true value of 12%.

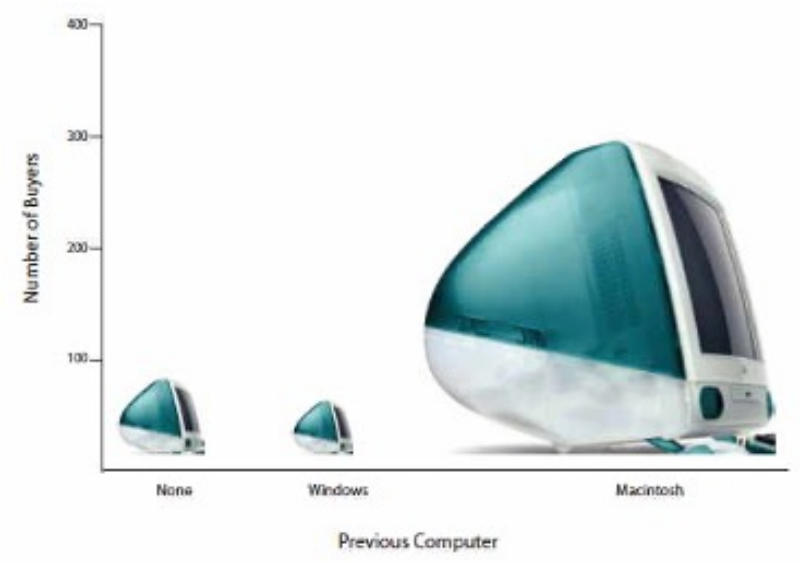


Figure 1.5: A redrawing of Figure 1.2 with a lie factor greater than 8.

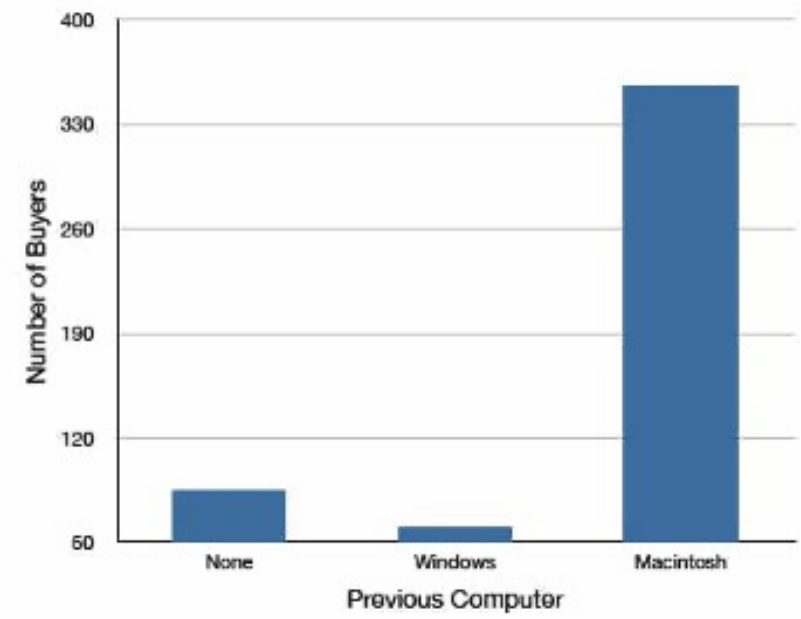


Figure 1.6: A redrawing of Figure 1.2 with a baseline of 50.

Finally, we note that it is a serious mistake to use a line graph when the X-axis contains merely qualitative variables. A line graph is essentially a bar graph with the tops of the bars represented by points joined by lines (the rest of the bar is suppressed). Figure 1.7 inappropriately shows a line graph of the card game data from Yahoo. The drawback to Figure 1.7 is that it gives the false impression that the games are naturally ordered in a numerical way when, in fact, they are ordered alphabetically.

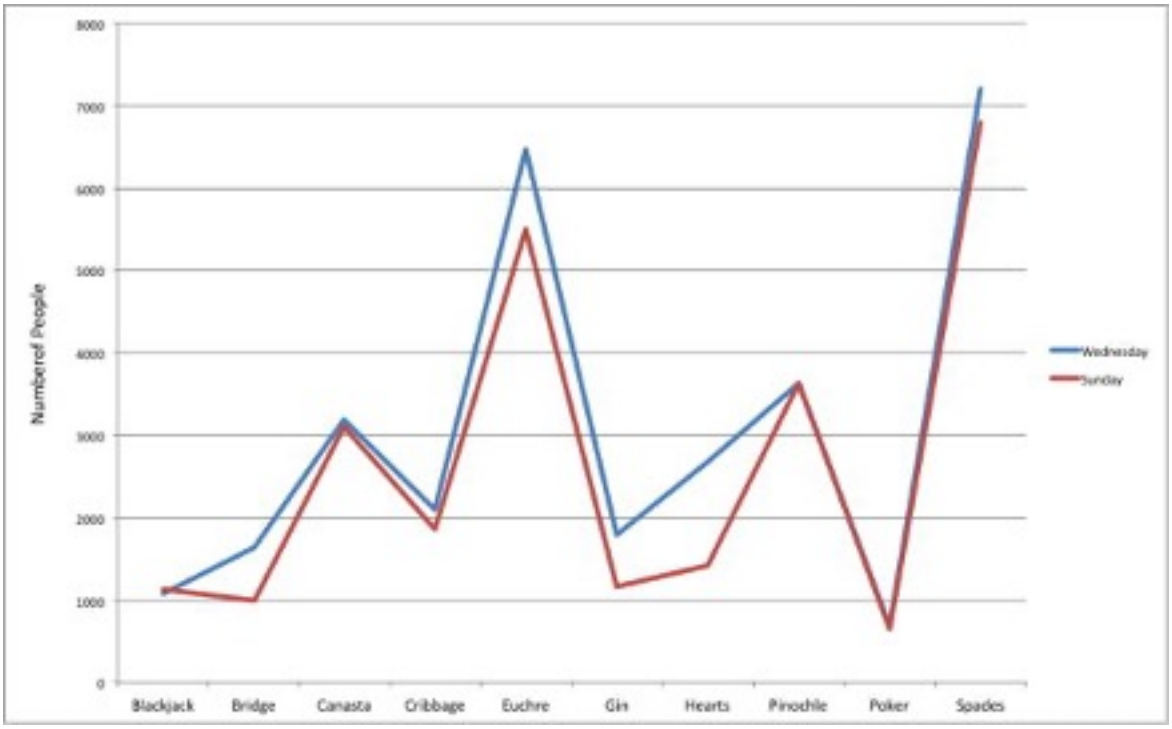


Figure 1.7: A line graph used inappropriately to depict the number of people playing different card games on Sunday and Wednesday.

1.4 Graphing Quantitative Variables

Having considered qualitative variables, we now turn our attention to some of the common types of graphs that are used to depict quantitative variables, beginning with histograms.

1.4.1 Histograms[4]

A histogram is a graphical method for displaying the shape of a distribution. It is particularly useful when there are a large number of observations. We begin with an example consisting of the scores of 642 students on a psychology test. The test consists of 197 items, each graded as "correct" or "incorrect." The students' scores ranged from 46 to 167.

The first step is to create a frequency table. Unfortunately, a simple frequency table would be too big, containing over 100 rows. To simplify the table, we group scores together as shown in Table 1.2.

Table 1.2: Grouped Frequency Distribution of Psychology Test Scores

Interval's Lower Limit	**Interval's Upper Limit**	**Class Frequency**
39.5	49.5	3
49.5	59.5	10
59.5	69.5	53
69.5	79.5	107
79.5	89.5	147
89.5	99.5	130
99.5	109.5	78
109.5	119.5	59
119.5	129.5	36
129.5	139.5	11
139.5	149.5	6
149.5	159.5	1
159.5	169.5	1

To create this table, the range of scores was broken into intervals, called class intervals or simply "bins." The first interval is from 39.5 to 49.5, the second from 49.5 to 59.5, etc. Next, the number of scores falling into each interval was counted to obtain the class frequencies. There are three scores in the first interval, 10 in the second, etc.

Class intervals of width 10 provide enough detail about the distribution to be revealing without making the graph too "choppy." More information on choosing the widths of class intervals is presented later in this section. Placing the limits of the class intervals midway between two numbers (e.g., 49.5) ensures that every score will fall in an interval rather than on the boundary between intervals.

In a histogram, the class frequencies are represented by bars. The height of each bar corresponds to its class frequency. A histogram of these data is shown in Figure 1.8.

[4]This section is adapted from David M. Lane. "Histograms." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/graphing_distributions/histograms.html

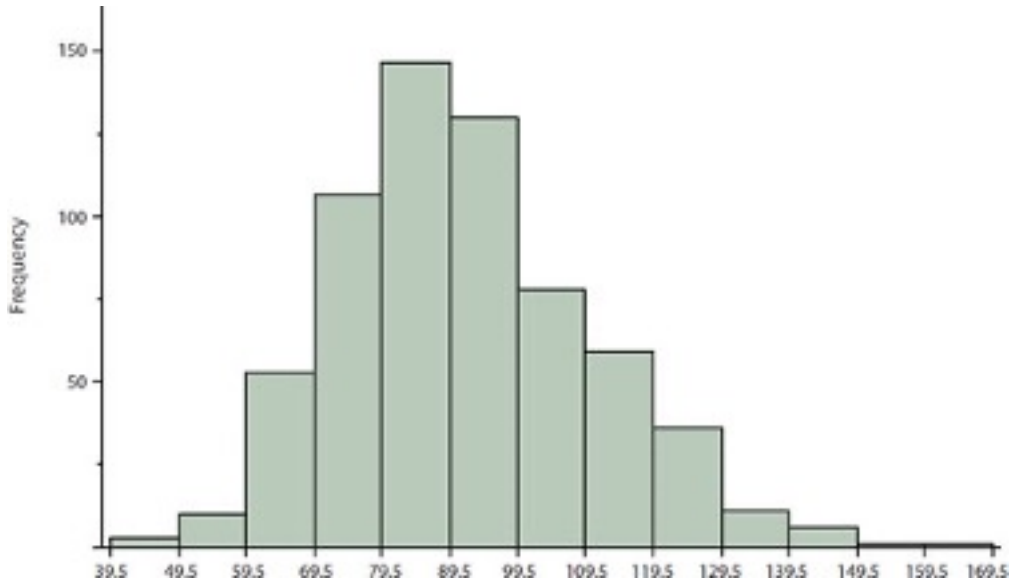


Figure 1.8: Histogram of scores on a psychology test.

The histogram makes it plain that most of the scores are in the middle of the distribution, with fewer scores in the extremes. You can also see that the distribution is not **symmetric**: the scores extend to the right farther than they do to the left. The distribution is therefore said to be **skewed**.

In our example, the observations are whole numbers. Histograms can also be used when the scores are measured on a more continuous scale such as the length of time (in milliseconds) required to perform a task. In this case, there is no need to worry about fence-sitters since they are improbable. (It would be quite a coincidence for a task to require exactly 7 seconds, measured to the nearest thousandth of a second.) We are therefore free to choose whole numbers as boundaries for our class intervals, for example, 4000, 5000, etc. The class frequency is then the number of observations that are greater than or equal to the lower bound, and strictly less than the upper bound. For example, one interval might hold times from 4000 to 4999 milliseconds. Using whole numbers as boundaries avoids a cluttered appearance, and is the practice of many computer programs that create histograms. Note also that some computer programs label the middle of each interval rather than the end points.

Histograms can be based on relative frequencies instead of actual frequencies. Histograms based on relative frequencies show the proportion of scores in each interval rather than the number of scores. In this case, the Y-axis runs from 0 to 1 (or somewhere in between if there are no extreme proportions). You can change a histogram based on frequencies to one based on relative frequencies by (a) dividing each class frequency by the total number of observations, and then (b) plotting the quotients on the Y-axis (labeled as proportion).

There is more to be said about the widths of the class intervals, sometimes called *bin widths*.

Your choice of bin width determines the number of class intervals. This decision, along with the choice of starting point for the first interval, affects the shape of the histogram. There are some "rules of thumb" that can help you choose an appropriate width. (But keep in mind that none of the rules is perfect.) We prefer the Rice rule, which is to set the number of intervals to twice the cube root of the number of observations. In the case of 1000 observations, the Rice rule yields 20 intervals. For the psychology test example used above, the Rice rule recommends 17. ***The best advice is to experiment with different choices of width, and to choose a histogram according to how well it communicates the shape of the distribution.***

1.4.2 Box Plots[5]

Box plots are useful for making comparisons and identifying **outliers**, meaning unusually large or small values for a variable. We will explain box plots with the help of data from an in-class experiment. As part of the "Stroop Interference Case Study,"[6] students in introductory statistics were presented with a page containing 30 colored rectangles. Their task was to name the colors as quickly as possible. Their times (in seconds) were recorded. We'll compare the scores for the 16 men and 31 women who participated in the experiment by making separate box plots for each gender. Such a display is said to involve *parallel box plots*.

There are several steps in constructing a box plot. The first relies on the 25^{th}, 50^{th}, and 75^{th} percentiles in the distribution of scores. Figure 1.9 shows how these three statistics are used. For each gender, we draw a box extending from the 25^{th} percentile to the 75^{th} percentile. The 50^{th} percentile is drawn inside the box. Therefore, the bottom of each box is the 25th percentile, the top is the 75th percentile, and the line in the middle is the 50th percentile. The data for the women in our sample are shown in Table 1.3.

Table 1.3: Women's times.

14	17	18	19	20	21	29
15	17	18	19	20	22	
16	17	18	19	20	23	
16	17	18	20	20	24	
17	18	18	20	21	24	

For these data, the 25th percentile is 17, the 50th percentile is 19, and the 75th percentile is 20. For the men (whose data are not shown), the 25th percentile is 19, the 50th percentile is 22.5, and the 75th percentile is 25.5.

Before proceeding, the terminology in Table 1.4 is helpful.

[5]This section is adapted from David M. Lane. "Box Plots." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/graphing_distributions/boxplots.html

[6]http://onlinestatbook.com/2/case_studies/stroop.html

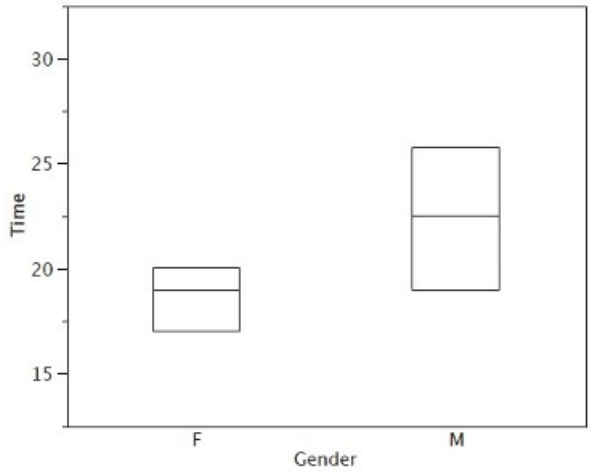
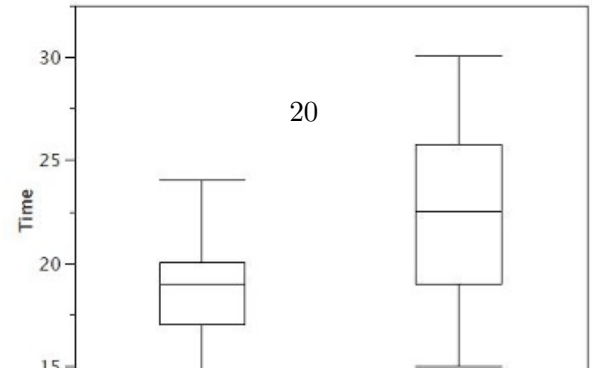


Figure 1.9: The first step in creating box plots.

Table 1.4: Box plot terms and values for women's times.

Name	Formula	Value
Upper Hinge	75th Percentile	20
Lower Hinge	25th Percentile	17
H-Spread	Upper Hinge - Lower Hinge	3
Step	1.5 x H-Spread	4.5
Upper Inner Fence	Upper Hinge + 1 Step	24.5
Lower Inner Fence	Lower Hinge - 1 Step	12.5
Upper Outer Fence	Upper Hinge + 2 Steps	29
Lower Outer Fence	Lower Hinge - 2 Steps	8
Upper Adjacent	Largest value below Upper Inner Fence	24
Lower Adjacent	Smallest value above Lower Inner Fence	14
Outside Value	A value beyond an Inner Fence but not beyond an Outer Fence	29
Far Out Value	A value beyond an Outer Fence	None

Continuing with the box plots, we put "whiskers" above and below each box to give additional information about the spread of the data. Whiskers are vertical lines that end in a horizontal stroke. Whiskers are drawn from the upper and lower hinges to the upper and lower adjacent values (24 and 14 for the women's data).



Although we don't draw whiskers all the way to outside or far out values, we still wish to represent them in our box plots. This is achieved by adding additional marks beyond the whiskers. Specifically, outside values are indicated by small "o's" and far out values are indicated by asterisks (*). In our data, there are no far out values and just one outside value. This outside value of 29 is for the women and is shown in Figure 1.11.

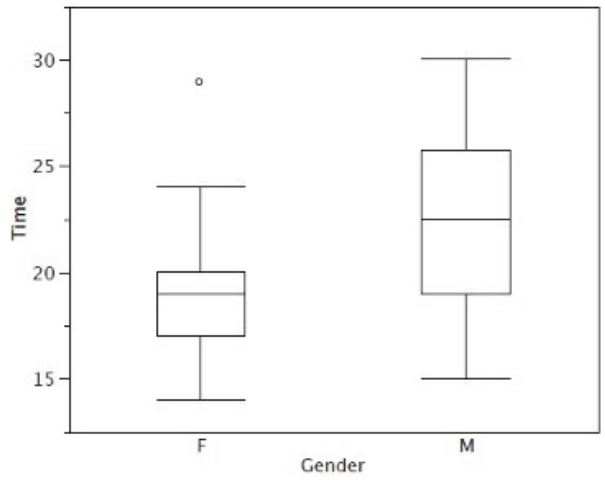


Figure 1.11: The box plots with the outside value shown.

There is one more mark to include in box plots (although sometimes it is omitted). We indicate the mean score for a group by inserting a plus sign. Figure 1.12 shows the result of adding means to our box plots.

Figure 1.12 provides a revealing summary of the data. Since half the scores in a distribution are between the hinges (recall that the hinges are the 25^{th} and 75^{th} percentiles), we see that half the women's times are between 17 and 20 seconds, whereas half the men's times are between 19 and 25.5. We also see that women generally named the colors faster than the men did, although one woman was slower than almost all of the men. Figure 1.13 shows the box plot for the women's data with detailed labels.

Box plots provide basic information about a distribution. For example, a distribution with a positive skew would have a longer whisker in the positive direction than in the negative direction. A larger mean than median would also indicate a positive skew. Box plots are good at portraying extreme values and are especially good at showing differences between distributions. However, many of the details of a distribution are not revealed in a box plot, and to examine these details one should create a histogram.

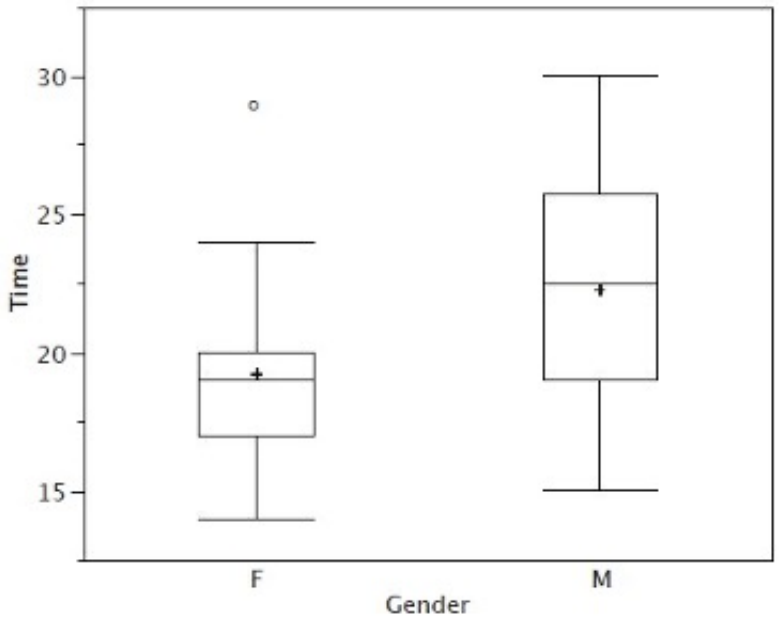


Figure 1.12: The completed box plots.

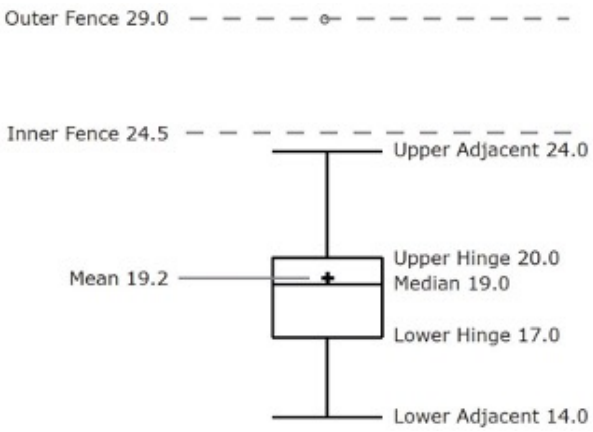


Figure 1.13: The box plot for the women's data with detailed labels.

1.4.3 Variations on box plots

Statistical analysis programs may offer options on how box plots are created. For example, the box plots in Figure 1.14 are constructed from our data but differ from the previous box plots in several ways.

1. It does not mark outliers.
2. The means are indicated by green lines rather than plus signs.
3. The mean of all scores is indicated by a gray line.
4. Individual scores are represented by dots. Since the scores have been rounded to the nearest second, any given dot might represent more than one score.
5. The box for the women is wider than the box for the men because the widths of the boxes are proportional to the number of subjects of each gender (31 women and 16 men).

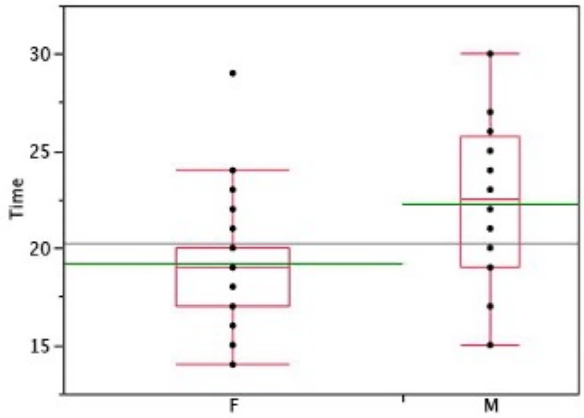


Figure 1.14: Box plots showing the individual scores and the means.

Each dot in Figure 1.14 represents a group of subjects with the same score (rounded to the nearest second). An alternative graphing technique is to "jitter" the points. This means spreading out different dots at the same horizontal position, one dot for each subject. The exact horizontal position of a dot is determined randomly (under the constraint that different dots don't overlap exactly). Spreading out the dots helps you to see multiple occurrences of a given score. However, depending on the dot size and the screen resolution, some points may be obscured even if the points are jittererd. Figure 1.15 shows what jittering looks like.

Different styles of box plots are best for different situations, and there are no firm rules for which to use. When exploring your data, you should try several ways of visualizing them. Which graphs you include in your report should depend on how well different graphs reveal the aspects of the data you consider most important.

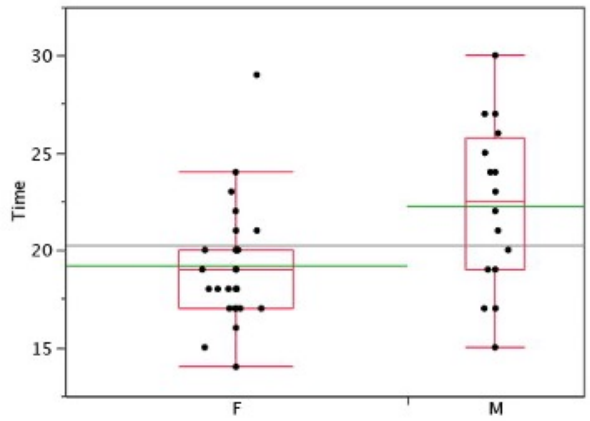


Figure 1.15: Box plots with the individual scores jittered.

1.4.4 Bar Charts for Quantitative Variables[7]

In the section on qualitative variables, we saw how bar charts could be used to illustrate the frequencies of different categories. For example, one bar chart showed how many purchasers of iMac computers were previous Macintosh users, previous Windows users, and new computer purchasers.

In this section, we show how bar charts can be used to present other kinds of quantitative information, not just frequency counts. The bar chart in Figure 1.16 shows the percent increases in the Dow Jones, Standard and Poor 500 (S & P), and Nasdaq stock indexes from May 24th 2000 to May 24th 2001. Notice that both the S & P and the Nasdaq had "negative increases" which means that they decreased in value. In this bar chart, the Y-axis is not frequency but rather the signed quantity percentage increase.

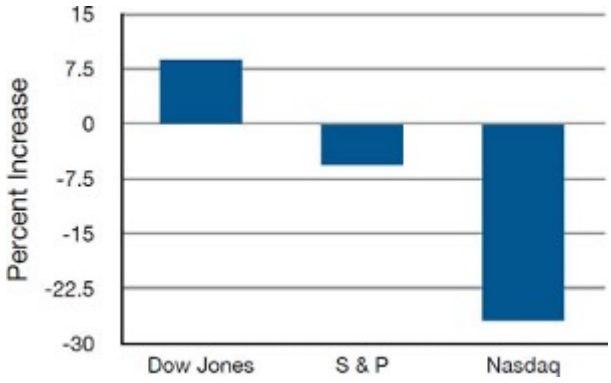


Figure 1.16: Percent increase in three stock indexes from May 24th 2000 to May 24th 2001.

[7]This section is adapted from David M. Lane. "Bar Charts." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/graphing_distributions/bar_chart.html

Bar charts are particularly effective for showing change over time. Figure 1.17, for example, shows the percent increase in the Consumer Price Index (CPI) over four three-month periods. The fluctuation in inflation is apparent in the graph.

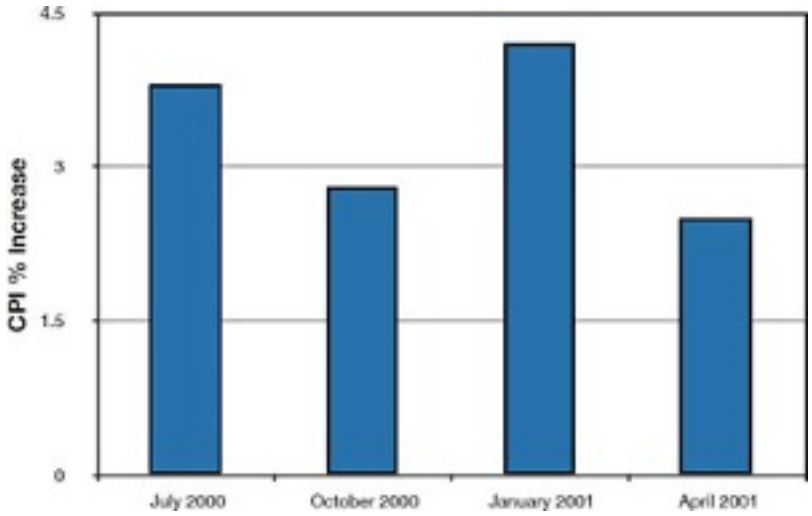


Figure 1.17: Percent change in the CPI over time. Each bar represents percent increase for the three months ending at the date indicated.

Bar charts are often used to compare the means of different experimental conditions. Figure 1.18 shows the mean time it took one of us (DL) to move the mouse to either a small target or a large target. On average, more time was required for small targets than for large ones.

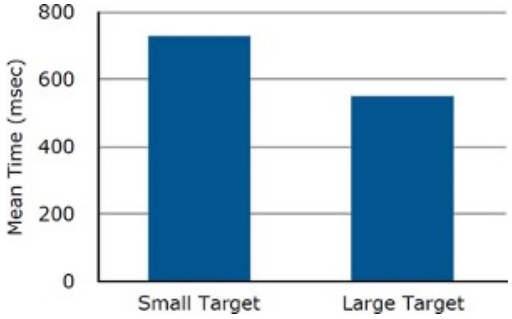


Figure 1.18: Bar chart showing the means for the two conditions.

Although bar charts can display means, we do not recommend them for this purpose. Box plots should be used instead since they provide more information than bar charts without taking up more space. For example, a box plot of the mouse-movement data is shown in Figure 1.19. You can see that Figure 1.19 reveals more about the distribution of movement times than does Figure 1.18.

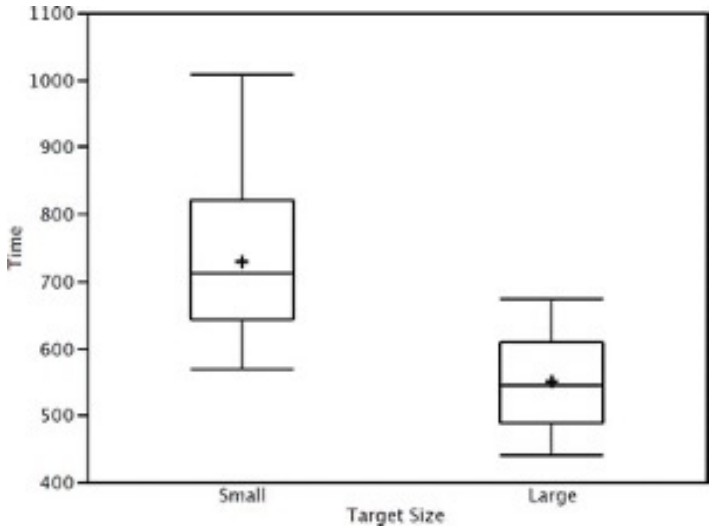


Figure 1.19: Box plots of times to move the mouse to the small and large targets.

The section on qualitative variables presented earlier in this chapter discussed the use of bar charts for comparing distributions. Some common graphical mistakes were also noted. The earlier discussion applies equally well to the use of bar charts to display quantitative variables.

Chapter 1 Appendix: Calculating Percentiles Under the Third Definition[8]

Let's begin with an example. Consider the 25th percentile for the 8 numbers in Table 1.5. Notice the numbers are given ranks ranging from 1 for the lowest number to 8 for the highest number.

Table 1.5: Test Scores.

Number	Rank

[8]This section is adapted from David M. Lane. "Percentiles." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/introduction/percentiles.html

3	1
5	2
7	3
8	4
9	5
11	6
13	7
15	8

The first step is to compute the rank (R) of the 25th percentile. This is done using the following formula:

$$R = P/100 \times (N + 1)$$

where P is the desired percentile (25 in this case) and N is the number of numbers (8 in this case). Therefore,

$$R = 25/100 \times (8 + 1) = 9/4 = 2.25.$$

If R is an integer, the Pth percentile is the number with rank R. When R is not an integer, we compute the Pth percentile by interpolation as follows:

1. Define IR as the integer portion of R (the number to the left of the decimal point). For this example, $IR = 2$.
2. Define FR as the fractional portion of R. For this example, $FR = 0.25$.
3. Find the scores with Rank IR and with Rank $IR + 1$. For this example, this means the score with Rank 2 and the score with Rank 3. The scores are 5 and 7.
4. Interpolate by multiplying the difference between the scores by FR and add the result to the lower score. For these data, this is (0.25)(7 - 5) + 5 = 5.5.

Therefore, the 25th percentile is 5.5. If we had used the first definition (the smallest score greater than 25% of the scores), the 25th percentile would have been 7. If we had used the second definition (the smallest score greater than or equal to 25% of the scores), the 25th percentile would have been 5.

For a second example, consider the 20 quiz scores shown in Table 1.6.

Table 1.6: 20 quiz scores.

Score	**Rank**
4	1
4	2
5	3
5	4
5	5
5	6
6	7
6	8
6	9
7	10
7	11
7	12
8	13
8	14
9	15
9	16
9	17
10	18
10	19
10	20

We will compute the 25th and the 85th percentiles. For the 25th,

$$R = 25/100 \times (20 + 1) = 21/4 = 5.25.$$

$$IR = 5 \text{ and } FR = 0.25.$$

Since the score with a rank of IR (which is 5) and the score with a rank of $IR + 1$ (which is 6) are both equal to 5, the 25th percentile is 5. In terms of the formula:

$$\text{25th percentile } = (.25) \times (5 - 5) + 5 = 5.$$

For the 85th percentile,

$$R = 85/100 \times (20 + 1) = 17.85.$$

$$IR = 17 \text{ and } FR = 0.85$$

Caution: FR does not generally equal the percentile to be computed as it does here.

The score with a rank of 17 is 9 and the score with a rank of 18 is 10. Therefore, the 85th percentile is:

$$(0.85)(10 - 9) + 9 = 9.85$$

Consider the 50th percentile of the numbers 2, 3, 5, 9.

$$R = 50/100 \times (4 + 1) = 2.5.$$

$$IR = 2 \text{ and } FR = 0.5.$$

The score with a rank of IR is 3 and the score with a rank of $IR + 1$ is 5. Therefore, the 50th percentile is:

$$(0.5)(5 - 3) + 3 = 4.$$

Finally, consider the 50th percentile of the numbers 2, 3, 5, 9, 11.

$$R = 50/100 \times (5 + 1) = 3.$$

$$IR = 3 \text{ and } FR = 0.$$

Whenever $FR = 0$, you simply find the number with rank IR. In this case, the third number is equal to 5, so the 50th percentile is 5. You will also get the right answer if you apply the general formula:

$$\text{50th percentile } = (0.00)(9 - 5) + 5 = 5.$$

2 Statistics for Describing One Variable at a Time

2.1 Measures of Central Tendency[1]

2.1.1 Mean

The **mean**[2] is the most common measure of central tendency. It is simply the sum of the numbers divided by the number of numbers. When using symbols and formulas to represent different statistics, we often distinguish between whether we are looking at a "sample" or a "population." We'll cover this distinction in more detail in our chapter about estimation. For now, think of a pollster who has conducted a survey with a sample of 1000 people. Even though only 1000 people responded to the survey, the pollster is actually interested in estimating the attitudes of a larger population—the entire public.

The symbol μ is used for the mean of a population. The symbol $\bar{X}$ is used for the mean of a sample. The formula for μ is shown below:

$$\mu = \Sigma X / N$$

where ΣX is the sum of all the numbers in the population and N is the number of numbers in the population.

The formula for $\bar{X}$ is essentially identical:

$$\bar{X} = \Sigma X / n$$

where ΣX is the sum of all the numbers in the sample and n is the number of numbers in the sample.

[1] This section is adapted from David M. Lane. "Measures of Central Tendency." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/summarizing_distributions/measures.html

[2] More specifically, the arithmetic mean is the most common measure of central tendency. Although the arithmetic mean is not the only "mean" (there is also a geometric mean), it is by far the most commonly used. Therefore, if the term "mean" is used without specifying whether it is the arithmetic mean, the geometric mean, or some other mean, it is assumed to refer to the arithmetic mean.

As an example, the mean of the numbers 1, 2, 3, 6, 8 is 20/5 = 4 regardless of whether the numbers constitute the entire population or just a sample from the population.

Table 2.1 shows the number of touchdown (TD) passes thrown by each of the 31 teams in the National Football League in the 2000 season. The mean number of touchdown passes thrown is 20.4516 as shown below.

$$\begin{aligned}\mu &= \Sigma X/N \\ &= 634/31 \\ &= 20.4516\end{aligned}$$

Table 2.1: Number of touchdown passes.

37 33 33 32 29 28 28 23 22 22 22 21 21 21 20 20 19 19 18 18 18 18 16 15 14 14 14 12 12 9 6

2.1.2 Median

The **median** is also a frequently used measure of central tendency. The median is the midpoint of a distribution: the same number of scores is above the median as below it. For the data in Table 2.1, there are 31 scores. The 16th highest score (which equals 20) is the median because there are 15 scores below the 16th score and 15 scores above the 16th score. The median can also be thought of as the 50th percentile.

2.1.2.1 Computation of the Median

When there is an odd number of numbers, the median is simply the middle number. For example, the median of 2, 4, and 7 is 4. When there is an even number of numbers, the median is the mean of the two middle numbers. Thus, the median of the numbers 2, 4, 7, 12 is (4+7)/2 = 5.5.

2.1.3 Mode

The **mode** is the most frequently occurring value. For the data in Table 2.1, the mode is 18 since more teams (4) had 18 touchdown passes than any other number of touchdown passes. With continuous data such as response time measured to many decimals, the frequency of each value is one since no two scores will be exactly the same (see discussion of continuous variables). Therefore the mode of continuous data is normally computed from a grouped frequency distribution. Table 2.2 shows a grouped frequency distribution for the target response

time data. Since the interval with the highest frequency is 600-700, the mode is the middle of that interval (650).

Table 2.2: Grouped frequency distribution.

Range	Frequency
500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

2.2 Comparing Measures of Central Tendency[3]

How do the various measures of central tendency compare with each other? For symmetric distributions, the mean and median are equal, as is the mode except in bimodal distributions. Differences among the measures occur with skewed distributions. Figure 2.1 shows the distribution of 642 scores on an introductory psychology test. The skew of this distribution can be described as slightly positive, meaning that there are more outliers on the positive (right) side of the distribution (see Section 5.1.4).

Measures of central tendency are shown in Table 2.3. Notice they do not differ greatly, with the exception that the mode is considerably lower than the other measures. When distributions have a positive skew, the mean is typically higher than the median, although it may not be in bimodal distributions. For these data, the mean of 91.58 is higher than the median of 90.

Table 2.3: Measures of central tendency for the test scores.

Measure	Value
Mode	84.00
Median	90.00
Mean	91.58

The distribution of baseball salaries (in 1994) shown in Figure 2.2 has a much more pronounced skew than the distribution in Figure 2.1.

Table 2.4 shows the measures of central tendency for these data. The large skew results in very different values for these measures. No single measure of central tendency is sufficient

[3]This section is adapted from David M. Lane. "Comparing Measures of Central Tendency." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/summarizing_distributions/comparing_measures.html

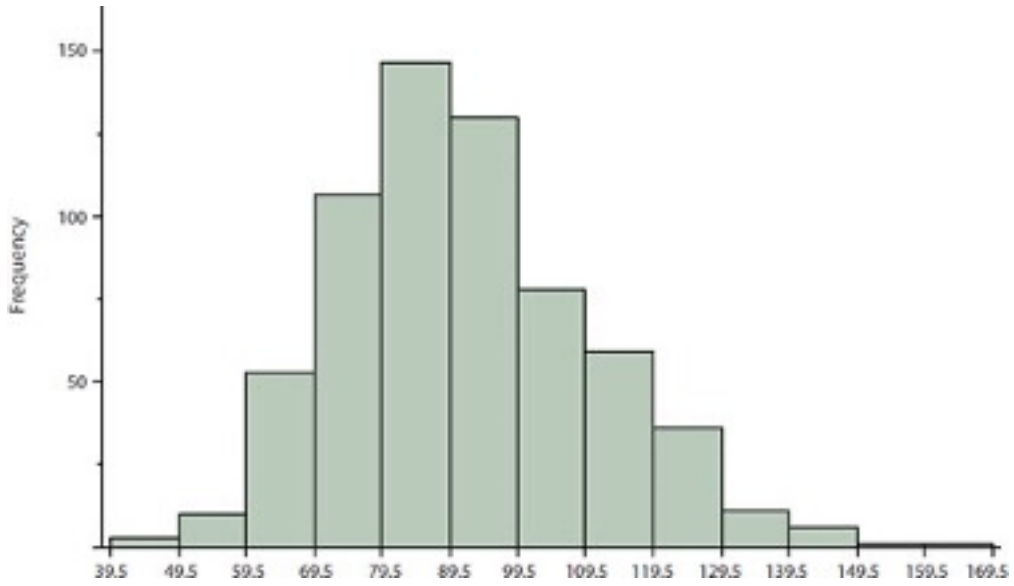


Figure 2.1: A distribution with a positive skew.

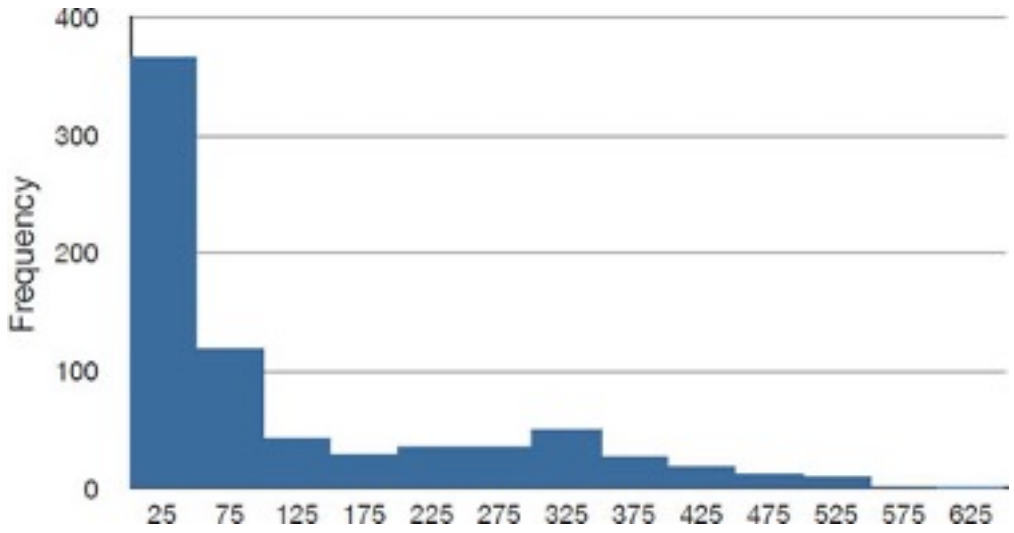


Figure 2.2: A distribution with a very large positive skew. This histogram shows the salaries of major league baseball players (in thousands of dollars: 25 equals 250,000).

for data such as these. If you were asked the very general question: "So, what do baseball players make?" and answered with the mean of $1,183,000, you would not have told the whole story since only about one third of baseball players make that much. If you answered with the mode of $250,000 or the median of $500,000, you would not be giving any indication that some players make many millions of dollars. Fortunately, there is no need to summarize a distribution with a single number. When the various measures differ, our opinion is that you should report the mean and the median. Sometimes it is worth reporting the mode as well. In the media, the median is usually reported to summarize the center of skewed distributions. You will hear about median salaries and median prices of houses sold, etc. This is better than reporting only the mean, but it would be informative to hear more statistics.

Table 2.4: Measures of central tendency for baseball salaries (in thousands of dollars).

Measure	**Value**
Mode	250
Median	500
Mean	1,183

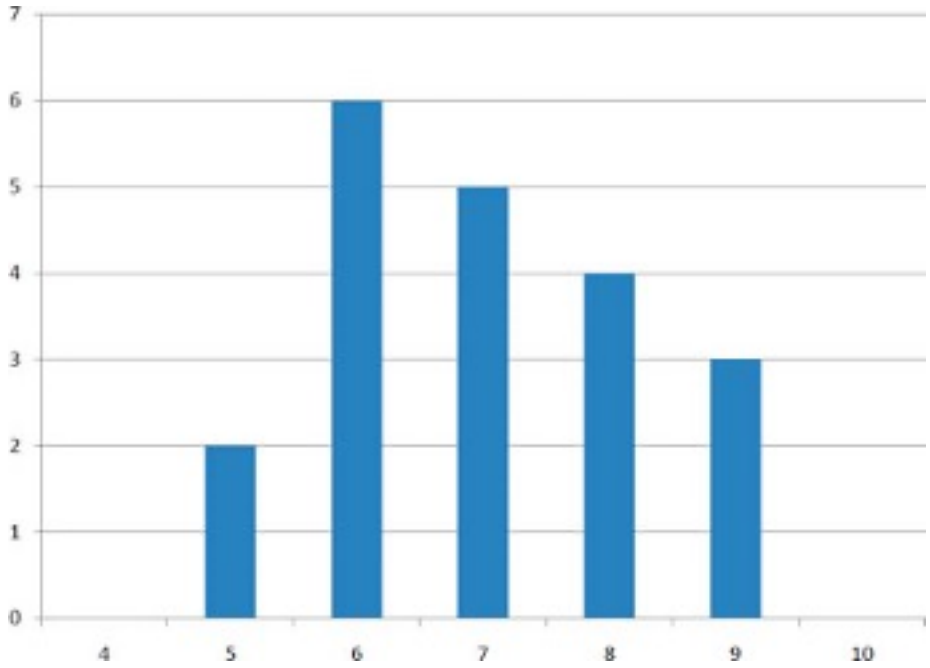
2.3 Measures of Spread[4]

2.3.1 What is Variability?

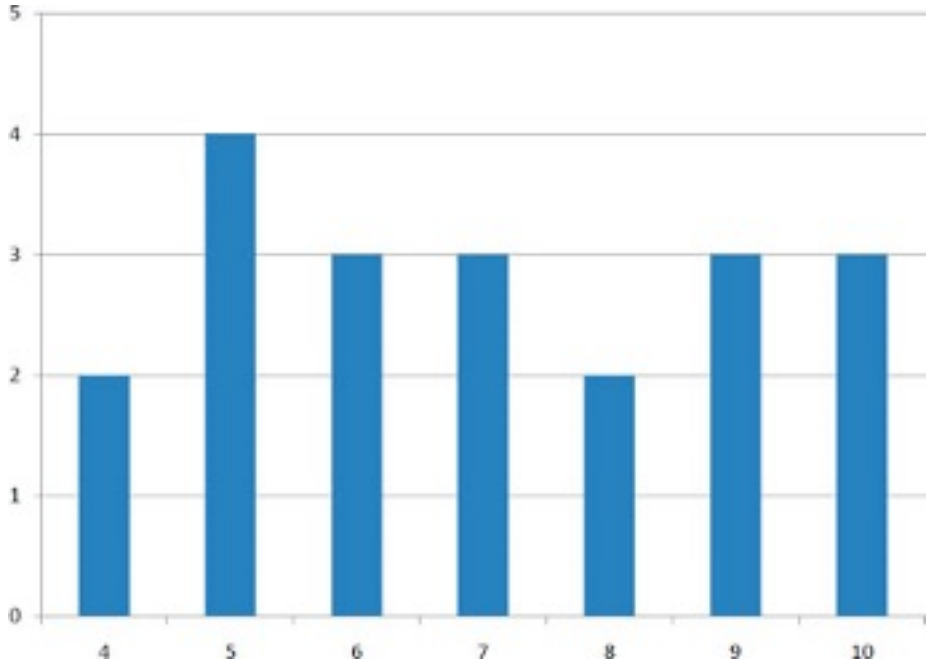
Variability refers to how "spread out" a group of scores is. To see what we mean by spread out, consider graphs in Figure 2.4. These graphs represent the scores on two quizzes. The mean score for each quiz is 7.0. Despite the equality of means, you can see that the distributions are quite different. Specifically, the scores on Quiz 1 are more densely packed and those on Quiz 2 are more spread out. The differences among students were much greater on Quiz 2 than on Quiz 1.

The terms variability, spread, and dispersion are synonyms, and refer to how spread out a distribution is. Just as in the section on central tendency where we discussed measures of the center of a distribution of scores, in this section we will discuss measures of the variability of a distribution. There are four frequently used measures of variability: the range, interquartile range, variance, and standard deviation. In the next few paragraphs, we will look at each of these four measures of variability in more detail.

[4]This section is adapted from David M. Lane. "Measures of Variability." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/summarizing_distributions/variability.html



(a) Quiz 1



(a) Quiz 2

Figure 2.4: Bar charts of two quizzes.

2.3.2 Range

The **range** is the simplest measure of variability to calculate, and one you have probably encountered many times in your life. The range is simply the highest score minus the lowest score. Let's take a few examples. What is the range of the following group of numbers: 10, 2, 5, 6, 7, 3, 4? Well, the highest number is 10, and the lowest number is 2, so 10 - 2 = 8. The range is 8. Let's take another example. Here's a dataset with 10 numbers: 99, 45, 23, 67, 45, 91, 82, 78, 62, 51. What is the range? The highest number is 99 and the lowest number is 23, so 99 - 23 equals 76; the range is 76. Now consider the two quizzes shown in Figure 2.4. On Quiz 1, the lowest score is 5 and the highest score is 9. Therefore, the range is 4. The range on Quiz 2 was larger: the lowest score was 4 and the highest score was 10. Therefore the range is 6.

2.3.3 Interquartile Range

The **interquartile range** (IQR) is the range of the middle 50% of the scores in a distribution. It is computed as follows:

$$IQR = \text{ 75th percentile } - \text{ 25th percentile}$$

For Quiz 1, the 75th percentile is 8 and the 25th percentile is 6. The interquartile range is therefore 2. For Quiz 2, which has greater spread, the 75th percentile is 9, the 25th percentile is 5, and the interquartile range is 4. Recall that in the discussion of box plots (Section 1.4.2), the 75th percentile was called the upper hinge and the 25th percentile was called the lower hinge. Thus, the interquartile range is neatly depicted by the box portion of a boxplot.

2.3.4 Variance

Variability can also be defined in terms of how close the scores in the distribution are to the middle of the distribution. Using the mean as the measure of the middle of the distribution, the **variance** is defined as the average squared difference of the scores from the mean. The data from Quiz 1 are shown in Table 2.5. The mean score is 7.0. Therefore, the column "Deviation from Mean" contains the score minus 7. The column "Squared Deviation" is simply the previous column squared.

Table 2.5: Calculation of Variance for Quiz 1 scores.

Scores	Deviation from Mean	Squared Deviation
9	2	4
9	2	4

9	2	4
8	1	1
8	1	1
8	1	1
8	1	1
7	0	0
7	0	0
7	0	0
7	0	0
7	0	0
6	-1	1
6	-1	1
6	-1	1
6	-1	1
6	-1	1
6	-1	1
5	-2	4
5	-2	4
	Means	
7	0	1.5

One thing that is important to notice is that the mean deviation from the mean is 0. This will always be the case. The mean of the squared deviations is 1.5. Therefore, the variance is 1.5. Analogous calculations with Quiz 2 show that its variance is 6.7. The formula for the variance is:

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

where $\sigma 2$ is the variance, μ is the mean, and N is the number of numbers. For Quiz 1, $\mu = 7$ and $N = 20$.

If the variance in a sample is used to estimate the variance in a population, then the previous formula underestimates the variance and the following formula should be used:

$$s^2 = \frac{\Sigma(X-\bar{X})^2}{n-1}$$

where s^2 is the estimate of the variance and $\bar{X}$ is the sample mean.

Note that $\bar{X}$ is the mean of a sample taken from a population with a mean of μ. Since, in practice, the variance is usually computed in a sample, this formula is most often used. While

it is not easy to succinctly explain why we divide by $n-1$ rather than simply n, the simulation "estimating variance"[5] illustrates the bias that arises if we use n as the denominator in the formula.

Let's look at a concrete example of calculating the sample variance. Assume the scores 1, 2, 4, and 5 were sampled from a larger population. To estimate the variance in the population you would compute s^2 as follows:

$$\bar{X} = (1 + 2 + 4 + 5)/4 = 12/4 = 3$$

$$s^2 = [(1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2]/(4-1) = (4 + 1 + 1 + 4)/3 = 10/3 = 3.333$$

2.3.5 Standard Deviation

The **standard deviation** is simply the square root of the variance. This makes the standard deviations of the two quiz distributions 1.225 and 2.588. We can interpret the standard deviation of X as approximating the typical distance between a given value of X and the mean of X. For example, suppose I tell you about a prison where the prisoners have a mean age of 42 years with a standard deviation of 8 years. If I randomly select one prisoner and ask you to guess their age, you should probably guess 42 since I've told you that is the mean. But even though 42 is your best guess, you can expect your guess to be off by about 8 years since the standard deviation is 8 (meaning the typical distance between a random prisoner's age and the mean age is approximately 8). You can't say ahead of time which direction your guess is likely to be off (guessing too old versus too young), just that you are likely to miss the reality for a randomly-selected individual by about 8 years on a typical guess (though any one guess may happen to be closer or further than 8 years).

2.4 Transforming Variables[6]

Often it is necessary to transform data from one measurement scale to another. For example, you might want to convert height measured in feet to height measured in inches. Table 2.6 shows the heights of four people measured in both feet and inches. To transform feet to inches, you simply multiply by 12. Similarly, to transform inches to feet, you divide by 12.

[5] https://onlinestatbook.com/2/summarizing_distributions/variance_est.html

[6] The initial part of this section is adapted from David M. Lane. "Linear Transformations." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/introduction/linear_transforms.html. There is also material adapted from David M. Lane. "Standard Normal Distribution." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/normal_distribution/standard_normal.html.

Table 2.6: Converting between feet and inches.

Feet	**Inches**
5.00	60
6.25	75
5.50	66
5.75	69

Some conversions require that you multiply by a number and then add a second number. A good example of this is the transformation between degrees Centigrade and degrees Fahrenheit. Table 2.7 shows the temperatures of 5 US cities in the early afternoon of November 16, 2002.

Table 2.7: Temperatures in 5 cities on 11/16/2002.

City	**Degrees Fahrenheit**	**Degrees Centigrade**
Houston	54	12.22
Chicago	37	2.78
Minneapolis	31	-0.56
Miami	78	25.56
Phoenix	70	21.11

The formula to transform Centigrade to Fahrenheit is:

$$F = 1.8C + 32$$

The formula for converting from Fahrenheit to Centigrade is

$$C = 0.5556F - 17.778$$

The transformation consists of multiplying by a constant and then adding a second constant. For the conversion from Centigrade to Fahrenheit, the first constant is 1.8 and the second is 32.

Figure 2.5 shows a plot of degrees Centigrade as a function of degrees Fahrenheit. Notice that the points form a straight line. This will always be the case if the transformation from one scale to another consists of multiplying by one constant and then adding a second constant. Such transformations are therefore called **linear transformations**.

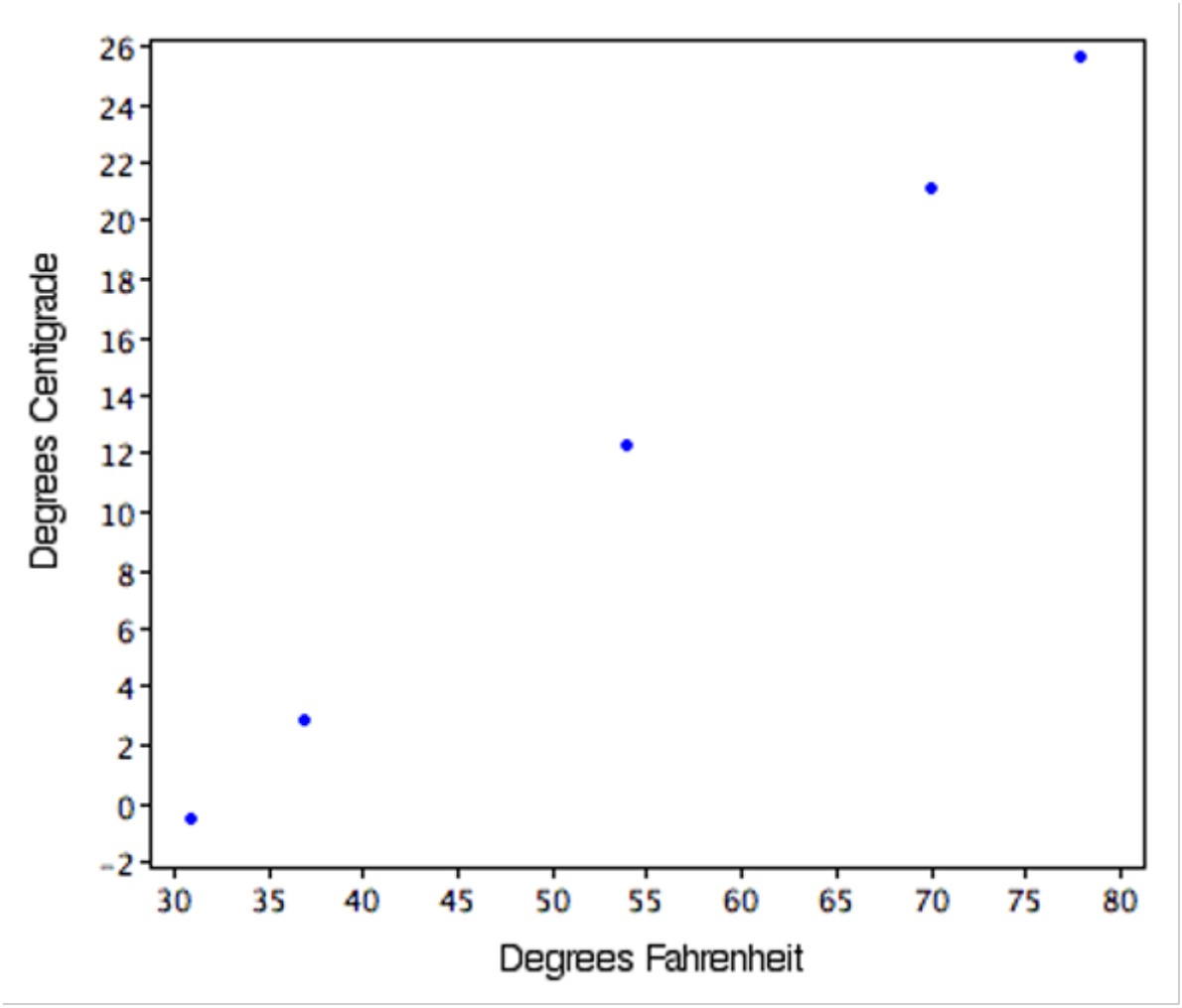


Figure 2.5: Degrees Centigrade as a function of degrees Fahrenheit.

2.4.1 Standardization (Z Scores)

So far, we've discussed transformations that are probably familiar to you. A type of transformation that may be new to you is **standardization** or creating Z scores. A value from any distribution can be transformed into a Z score using the following formula:

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

where Z is the new value, X is the value on the original distribution, μ is the mean of the original distribution, and σ is the standard deviation of the original distribution.

As a simple application, suppose you want the Z score for a value of 26 taken from a distribution with a mean of 50 and a standard deviation of 10. Applying the formula, we obtain:

$$Z = (26 - 50)/10 = -2.4$$

If all the values in a distribution are transformed to Z scores, then the new distribution will have a mean of 0 and a standard deviation of 1. This process of transforming a distribution to one with a mean of 0 and a standard deviation of 1 is called standardizing the distribution. Sometimes it will be easier to work with a standardized version of a variable.

2.4.2 Log Transformations[7]

Sometimes it is also useful to use transformations that are not linear. For example, the log transformation can be used to make highly skewed distributions less skewed. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics (see Chapter 4).

Figure 2.6 shows an example of how a log transformation can make patterns more visible. Both graphs plot the brain weight of animals as a function of their body weight. The raw weights are shown in the upper panel; the log-transformed weights are plotted in the lower panel.

It is hard to discern a pattern in the upper panel whereas the strong relationship is shown clearly in the lower panel.

[7]This subsection is adapted from David M. Lane. "Log Transformations." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/transformations/log.html

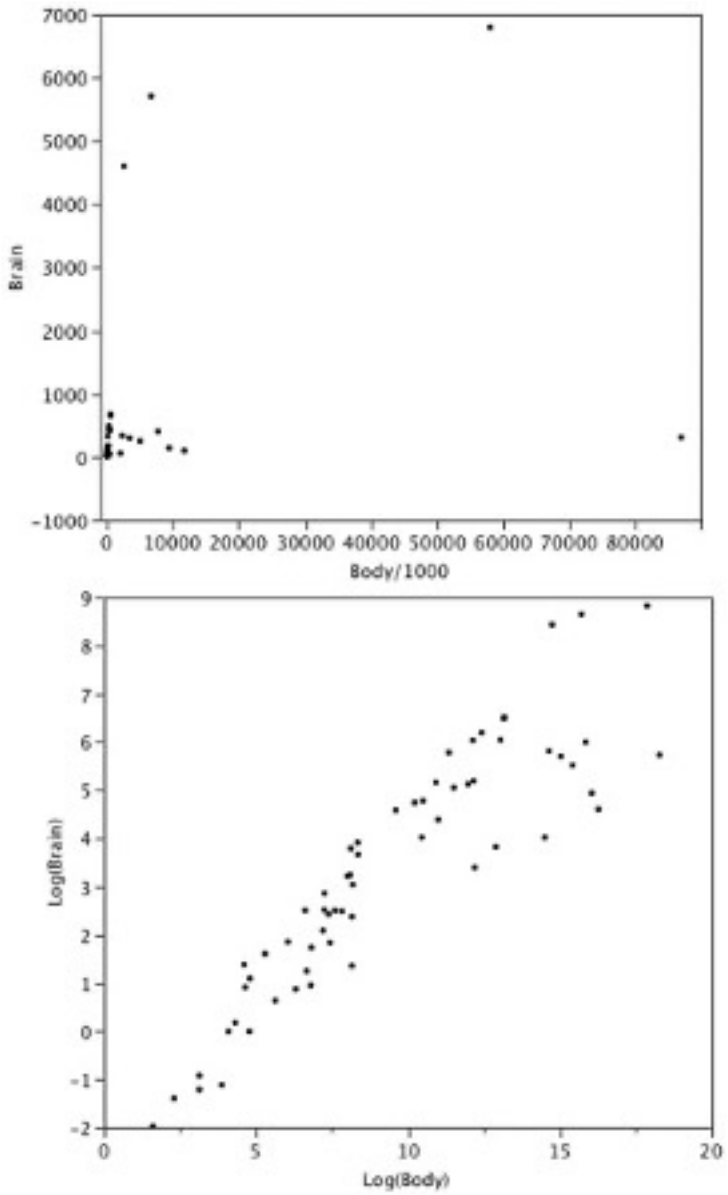


Figure 2.6: Scatter plots of brain weight as a function of body weight in terms of both raw data (upper panel) and log-transformed data (lower panel).

3 Tools for Describing the Relationship Between Two Quantitative Variables

3.1 Introduction to Bivariate Data[1]

Measures of central tendency, variability, and spread summarize a single variable by providing important information about its distribution. Often, more than one variable is collected on each individual. For example, in large health studies of populations it is common to obtain variables such as age, sex, height, weight, blood pressure, and total cholesterol on each individual. Economic studies may be interested in, among other things, personal income and years of education. As a third example, most university admissions committees ask for an applicant's high school grade point average and standardized admission test scores (e.g., SAT). In this chapter we consider bivariate data, which for now consists of two quantitative variables for each individual. Our first interest is in summarizing such data in a way that is analogous to summarizing univariate (single variable) data.

By way of illustration, let's consider something with which we are all familiar: age. Let's begin by asking if people tend to marry other people of about the same age. Our experience tells us "yes," but how good is the correspondence? One way to address the question is to look at pairs of ages for a sample of married couples. Table 3.1 below shows the ages of 10 married couples. Going across the columns we see that, yes, husbands and wives tend to be of about the same age, with men having a tendency to be slightly older than their wives. This is no big surprise, but at least the data bear out our experiences, which is not always the case.

Table 3.1: Sample of spousal ages of 10 White American Couples.

Husband	36	72	37	36	51	50	47	50	37	41
Wife	35	67	33	35	50	46	47	42	36	41

The pairs of ages in Table 3.1 are from a dataset consisting of 282 pairs of spousal ages, too many to make sense of from a table. What we need is a way to summarize the 282 pairs of ages. We know that each variable can be summarized by a histogram (see Figure 3.1) and by a mean and standard deviation (See Table 3.2).

[1]This section is adapted from Rudy Guerra and David M. Lane. "Introduction to Bivariate Data." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/describing_bivariate_data/intro.html

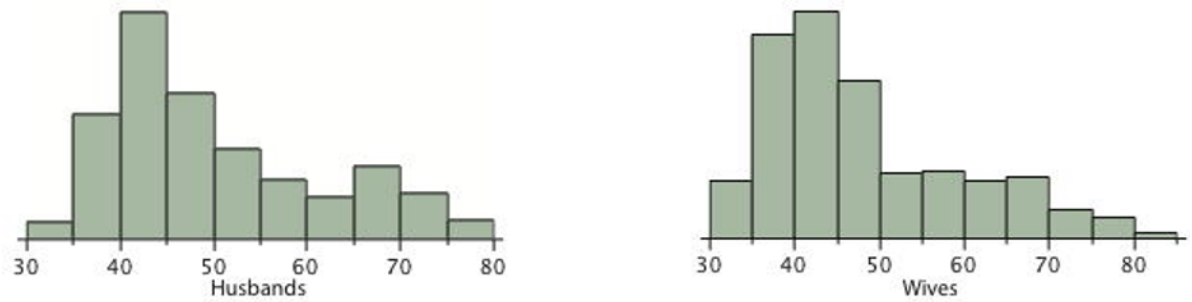


Figure 3.1: Histograms of spousal ages.

Table 3.2: Means and standard deviations of spousal ages.

	Mean	**Standard Deviation**
Husbands	49	11
Wives	47	11

Each distribution is fairly skewed with a long right tail. From Table 3.1 we see that not all husbands are older than their wives and it is important to see that this fact is lost when we separate the variables. That is, even though we provide summary statistics on each variable, the pairing within couple is lost by separating the variables. We cannot say, for example, based on the means alone what percentage of couples has younger husbands than wives. We have to count across pairs to find this out. Only by maintaining the pairing can meaningful answers be found about couples per se. Another example of information not available from the separate descriptions of husbands and wives' ages is the mean age of husbands with wives of a certain age. For instance, what is the average age of husbands with 45-year-old wives? Finally, we do not know the relationship between the husband's age and the wife's age.

We can learn much more by displaying the bivariate data in a graphical form that maintains the pairing. Figure 3.2 shows a scatter plot of the paired ages. The x-axis represents the age of the husband and the y-axis the age of the wife.

There are two important characteristics of the data revealed by Figure 3.2. First, it is clear that there is a strong relationship between the husband's age and the wife's age: the older the husband, the older the wife. When one variable (Y) increases with the second variable (X), we say that X and Y have a **positive association**. Conversely, when Y decreases as X increases, we say that they have a **negative association**.

Second, the points cluster along a straight line. When this occurs, the relationship is called a **linear relationship**.

Figure 3.3 shows a scatter plot of Arm Strength and Grip Strength from 149 individuals working in physically demanding jobs including electricians, construction and maintenance workers, and auto mechanics. Not surprisingly, the stronger someone's grip, the stronger their arm tends to be. There is therefore a positive association between these variables. Although the points cluster along a line, they are not clustered quite as closely as they are for the scatter plot of spousal age.

Not all scatter plots show linear relationships. Figure 3.4 shows the results of an experiment conducted by Galileo on projectile motion.[2] In the experiment, Galileo rolled balls down an incline and measured how far they traveled as a function of the release height. It is clear from Figure 3.4 that the relationship between "Release Height" and "Distance Traveled" is

[2] https://www.amstat.org/publications/jse/v3n1/datasets.dickey.html

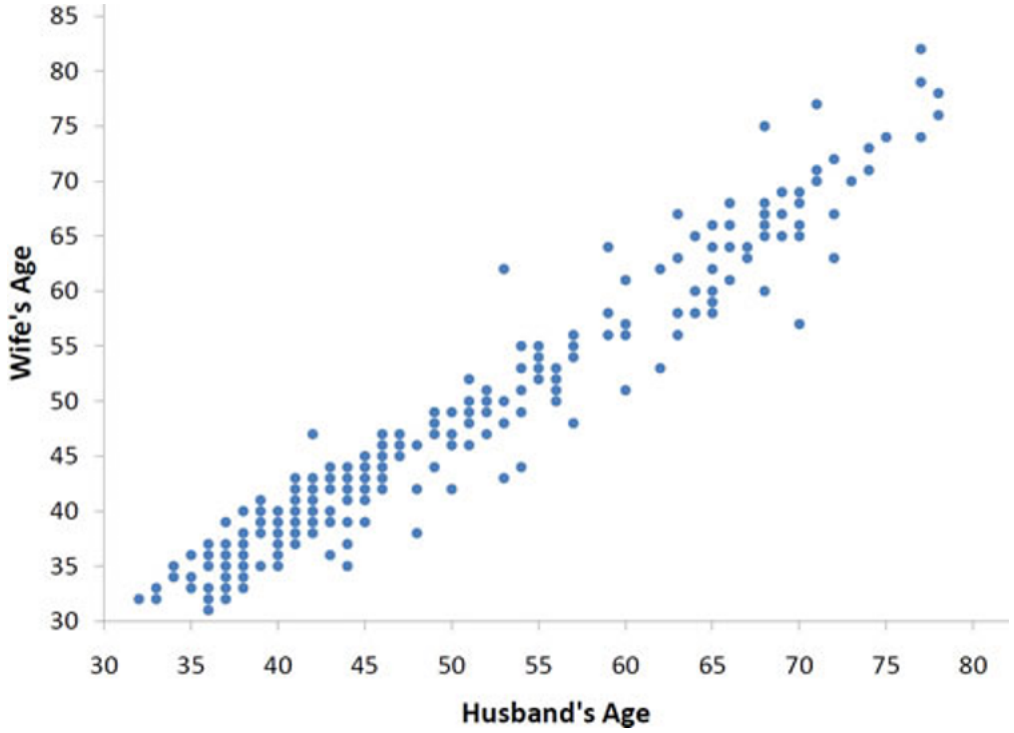


Figure 3.2: Scatter plot showing wife's age as a function of husband's age, r = 0.97.

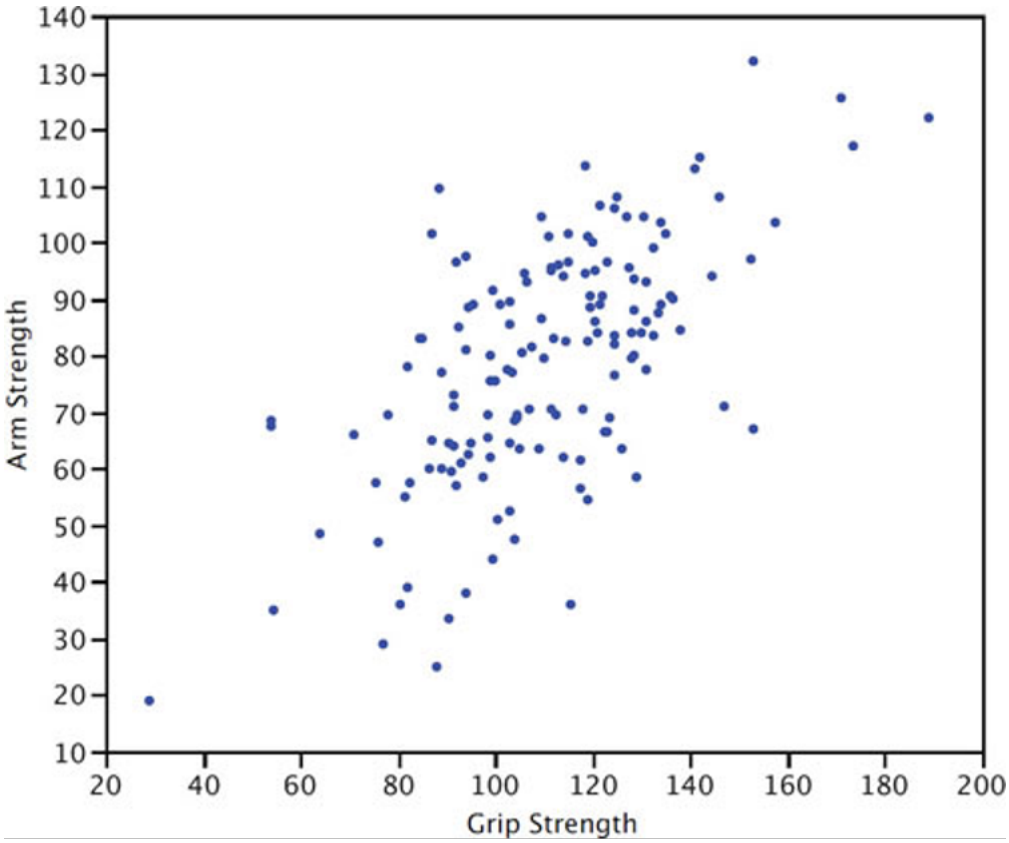


Figure 3.3: Scatter plot of Grip Strength and Arm Strength, r = 0.63.

not described well by a straight line: If you drew a line connecting the lowest point and the highest point, all of the remaining points would be above the line. The data are better fit by a parabola (a type of curved line).

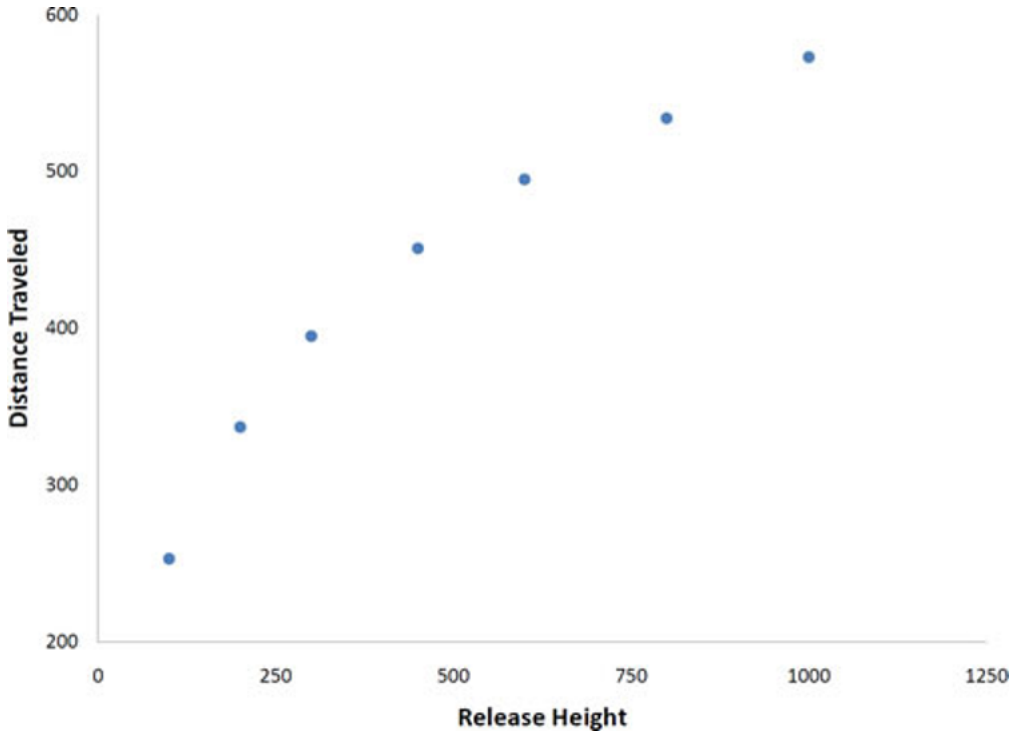


Figure 3.4: Galileo's data showing a non-linear relationship.

Scatter plots that show linear relationships between variables can differ in several ways including the slope of the line about which they cluster and how tightly the points cluster about the line. We now turn our attention to a statistical measure of the strength of the relationship between two quantitative variables.

3.2 What is Correlation?[3]

The Pearson product-moment correlation coefficient is a measure of the strength of the linear relationship between two variables. It is referred to as Pearson's correlation or simply as the **correlation coefficient**. If the relationship between the variables is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables.

[3]This section is adapted from David M. Lane. "Values of the Pearson Correlation." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/describing_bivariate_data/pearson.html

The symbol for Pearson's correlation is "ρ" when it is measured in the population and "r" when it is measured in a sample. Because we will be dealing almost exclusively with samples, we will use r to represent Pearson's correlation unless otherwise noted.

Pearson's r can range from -1 to 1. An r of -1 indicates a perfect negative linear relationship between variables, an r of 0 indicates no linear relationship between variables, and an r of 1 indicates a perfect positive linear relationship between variables. Figure 3.5 shows a scatter plot for which r = 1.

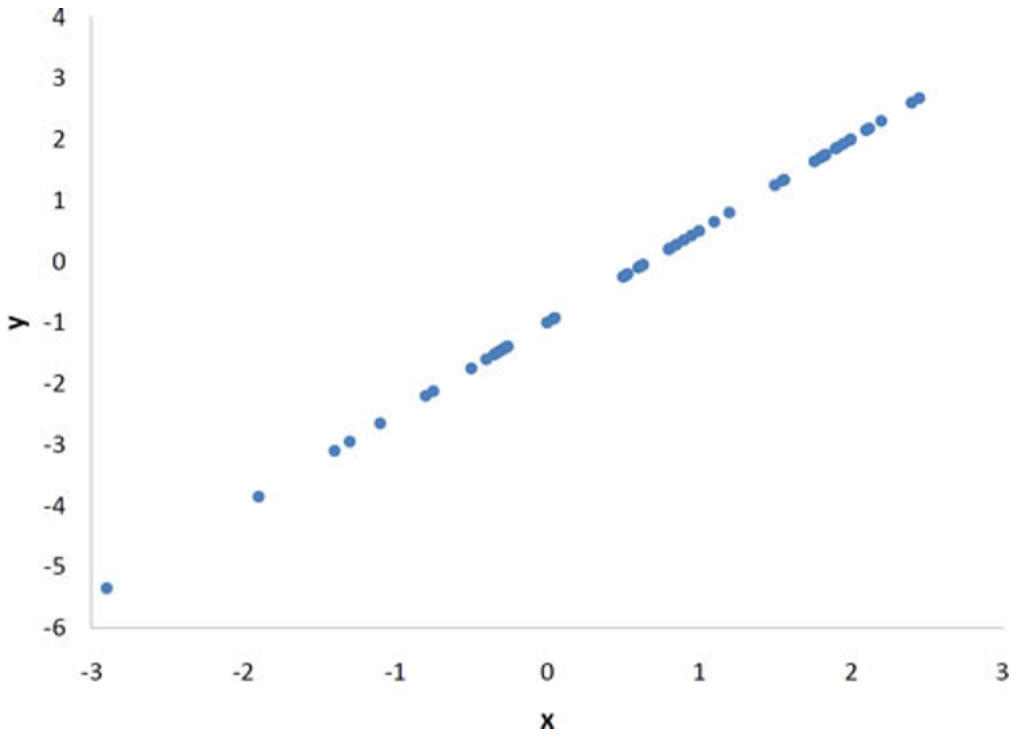


Figure 3.5: A perfect positive linear relationship, r = 1.

With real data, you would not expect to get values of r of exactly -1, 0, or 1. The data for spousal ages shown earlier in this chapter in Figure 3.2 has an r of 0.97.

The relationship between grip strength and arm strength depicted in Figure 3.3 (also described in the introductory section) is 0.63.

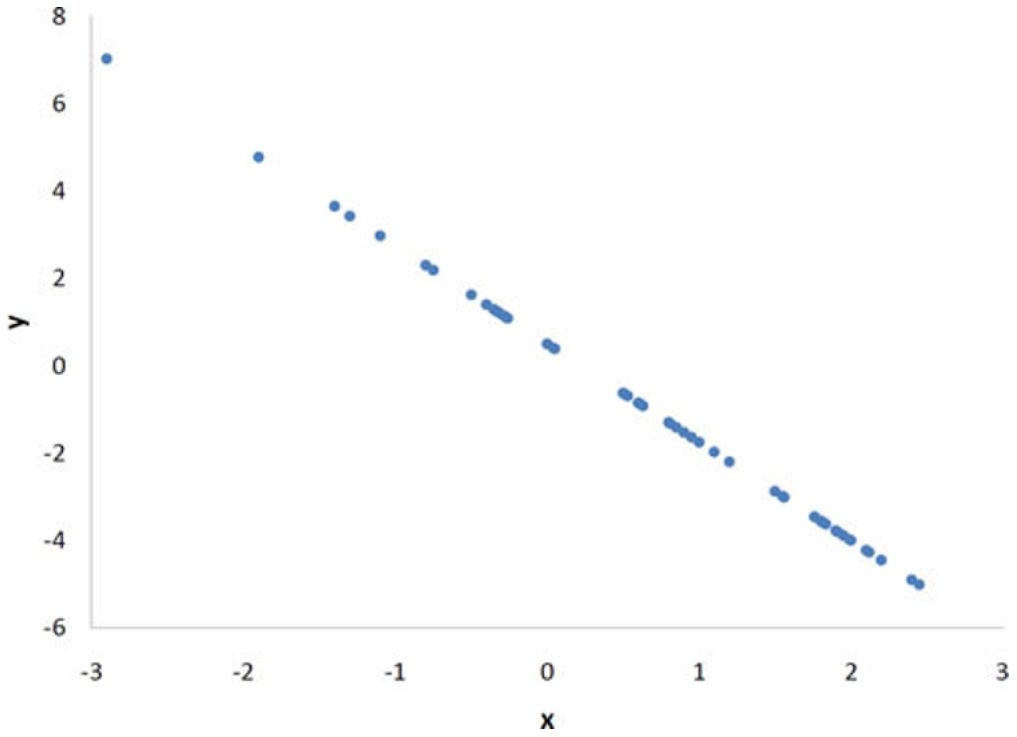


Figure 3.6: A perfect negative linear relationship, r = 1.

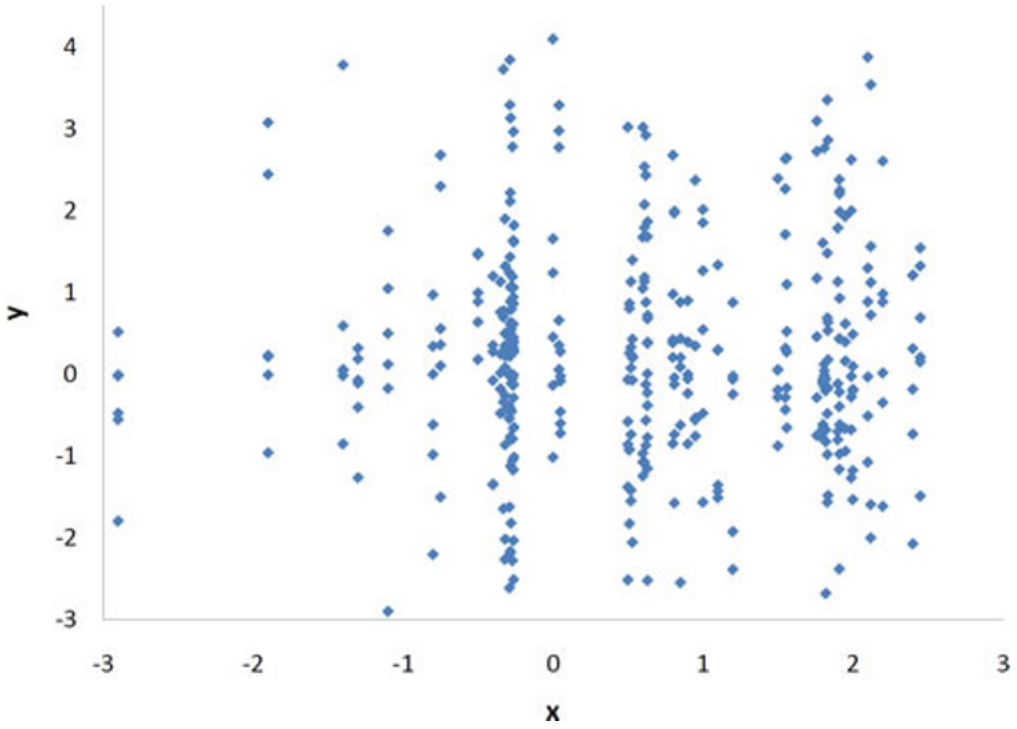


Figure 3.7: A scatter plot for which r = 0. Notice that there is no relationship between X and Y.

3.3 How Correlation is Calculated[4]

There are several formulas that can be used to compute Pearson's correlation. Some formulas make more conceptual sense whereas others are easier to actually compute. We are going to begin with a formula that makes more conceptual sense.

We are going to compute the correlation between the variables X and Y shown in Table 3.3. We begin by computing the mean for X and subtracting this mean from all values of X. The new variable is called "x." The variable "y" is computed similarly. The variables x and y are said to be deviation scores because each score is a deviation from the mean. Notice that the means of x and y are both 0. Next we create a new column by multiplying x and y.

Before proceeding with the calculations, let's consider why the sum of the xy column reveals the relationship between X and Y. If there were no relationship between X and Y, then positive values of x would be just as likely to be paired with negative values of y as with positive values. This would make negative values of xy as likely as positive values and the sum would be small. On the other hand, consider Table 3.3 in which high values of X are associated with high values of Y and low values of X are associated with low values of Y. You can see that positive values of x are associated with positive values of y and negative values of x are associated with negative values of y. In all cases, the product of x and y is positive, resulting in a high total for the xy column. Finally, if there were a negative relationship then positive values of x would be associated with negative values of y and negative values of x would be associated with positive values of y. This would lead to negative values for xy.

Table 3.3: Calculation of r.

	X	**Y**	**x**	**y**	**xy**	**x^2**	**y^2**
	1	4	-3	-5	15	9	25
	3	6	-1	-3	3	1	9
	5	10	1	1	1	1	1
	5	12	1	3	3	1	9
	6	13	2	4	8	4	16
Total	20	45	0	0	30	16	60
Mean	4	9	0	0	6		

Pearson's r is designed so that the correlation between height and weight is the same whether height is measured in inches or in feet. To achieve this property, Pearson's correlation is computed by dividing the sum of the xy column ($\sum xy$) by the square root of the product of the sum of the x^2 column ($\sum x^2$) and the sum of the y^2 column ($\sum y^2$). The resulting formula is:

[4]This section is adapted from David M. Lane. "Computing Pearson's r." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/describing_bivariate_data/calculation.html

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

and therefore:

$$r = \frac{30}{\sqrt{(16)(60)}} = \frac{30}{\sqrt{960}} = \frac{30}{30.984} = 0.968$$

An alternative computational formula that avoids the step of computing deviation scores is:

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{n}\right)}\sqrt{\left(\sum Y^2 - \frac{(\sum Y)^2}{n}\right)}}$$

3.4 Introduction to Linear Regression[5]

In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the **dependent variable** and is referred to as Y. The variable we are basing our predictions on is called the **independent variable** and is referred to as X. When there is only one predictor variable, the prediction method is called simple regression. In simple linear regression, the topic of this section, the predictions of Y when plotted as a function of X form a straight line.

The example data in Table 3.4 are plotted in Figure 3.8. You can see that there is a positive relationship between X and Y. If you were going to predict Y from X, the higher the value of X, the higher your prediction of Y.

Table 3.4: Example data.

X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line. The black diagonal line in Figure 3.9 is the regression line

[5]This section is adapted from David M. Lane. "Introduction to Linear Regression." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/regression/intro.html

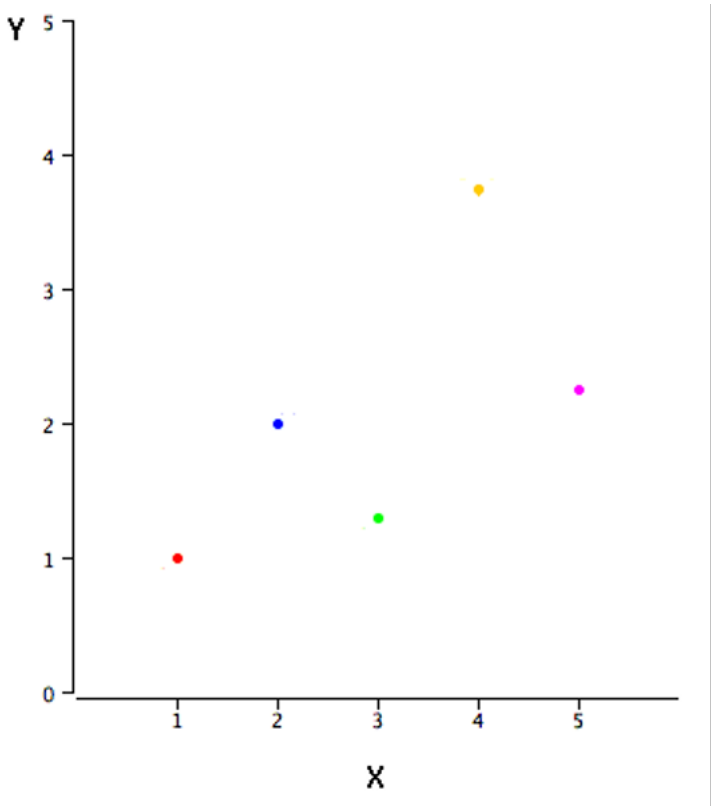


Figure 3.8: A scatter plot of the example data.

and consists of the predicted score on Y for each possible value of X. The vertical lines from the points to the regression line represent the errors of prediction. As you can see, the red point is very near the regression line; its error of prediction is small. By contrast, the yellow point is much higher than the regression line and therefore its error of prediction is large.

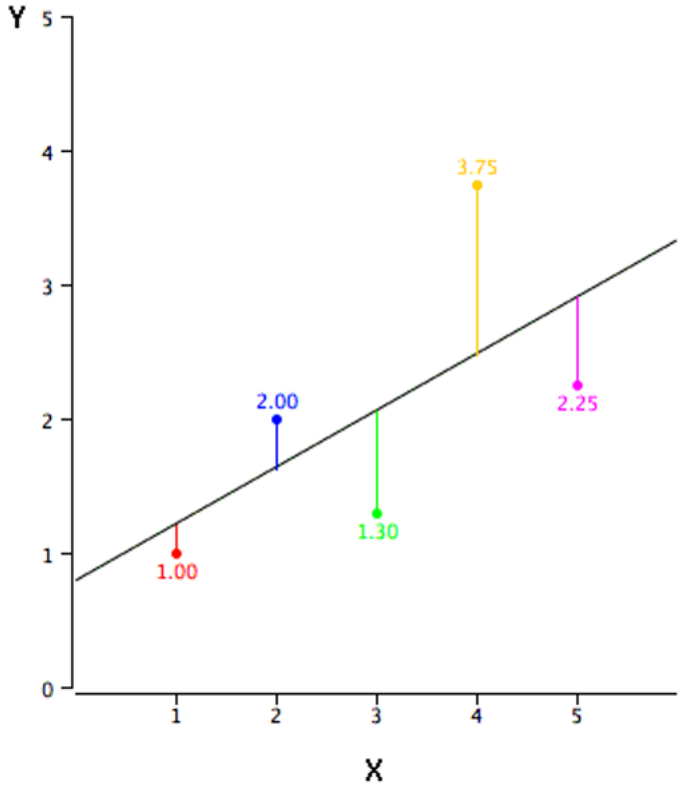


Figure 3.9: A scatter plot of the example data. The black line consists of the predictions, the points are the actual data, and the vertical lines between the points and the black line represent errors of prediction.

The error of prediction for a point is the value of the point minus the predicted value (the value on the line). Table 3.5 shows the predicted values ($\hat{Y}$) and the errors of prediction ($Y - \hat{Y}$). For example, the first point has a Y of 1.00 and a predicted Y (called $\hat{Y}$) of 1.21. Therefore, its error of prediction is -0.21.

Table 3.5: Example data.

X	Y	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
1.00	1.00	1.210	-0.210	0.044
2.00	2.00	1.635	0.365	0.133
3.00	1.30	2.060	-0.760	0.578

X	Y	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
4.00	3.75	2.485	1.265	1.600
5.00	2.25	2.910	-0.660	0.436

You may have noticed that we did not specify what is meant by "best-fitting line." By far, the most commonly-used criterion for the best-fitting line is the line that minimizes the sum of the squared errors of prediction. That is the criterion that was used to find the line in Figure 3.9. The last column in Table 3.5 shows the squared errors of prediction. The sum of the squared errors of prediction shown in Table 3.5 is lower than it would be for any other regression line.

The formula for a regression line is

$$\hat{Y} = \alpha + \beta X$$

where $\hat{Y}$ is the predicted score, α is the Y-intercept, and β is the slope of the line. The equation for the line in Figure 3.9 is

$$\hat{Y} = 0.785 + 0.425X$$

Using this equation, we can calculate predictions for Y based on the value of X. For $X = 1$,

$$\hat{Y} = 0.785 + (0.425)(1) = 1.21.$$

For $X = 2$,

$$\hat{Y} = 0.785 + (0.425)(2) = 1.64.$$

3.4.1 A Real Example

The case study "SAT and College GPA"[6] contains high school and university grades for 105 computer science majors at a local state school. We now consider how we could predict a student's university GPA if we knew his or her high school GPA.

Figure 3.10 shows a scatter plot of University GPA as a function of High School GPA. You can see from the figure that there is a strong positive relationship. The correlation is 0.78. The regression equation is

[6] http://onlinestatbook.com/2/case_studies/sat.html

$$\widehat{University\_GPA} = (0.675)(HighSchoolGPA) + 1.097$$

Therefore, a student with a high school GPA of 3 would be predicted to have a university GPA of

$$\widehat{University\_GPA} = (0.675)(3) + 1.097 = 3.12.$$

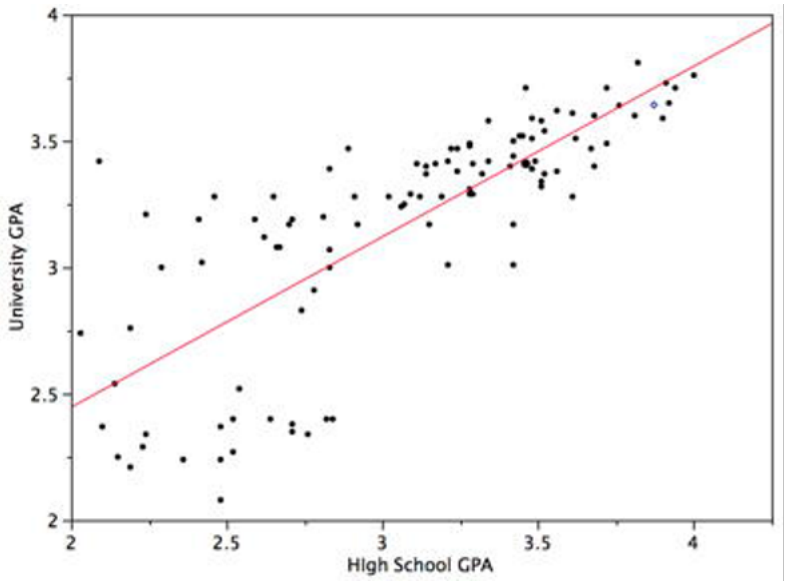


Figure 3.10: University GPA as a function of High School GPA.

3.5 Quick Guide to Interpreting Regression Results[7]

Many social science papers report their main results in the form of a regression table. It's fairly easy to get started interpreting these results using the three S's:[8]

- **Significance**: Is the relationship between the two variables strong enough (relative to the precision of the estimate) to be considered statistically reliable?[9] To assess this, check the p-value. For now, you can use the following rule-of-thumb:

[7]This section is written by Nathan Favero.

[8]Wheelan, C. (2010.) *Introduction to Public Policy.* New York: W. W. Norton & Company.

[9]In other words, can we conclude it is signal rather than noise? See: Fricker Jr, R. D., Burke, K., Han, X., & Woodall, W. H. (2019). Assessing the statistical analyses used in basic and applied social psychology after their p-value ban. *The American Statistician*, *73*(sup1), 374-384.

  - If $p < 0.05$: the relationship is statistically significant; proceed to evaluating sign and size.
  - If $p > 0.05$: results are somewhat indeterminate; any association detected between the two variables could easily be caused by coincidence or random "noise" (so you may want to skip evaluating sign and size)

- **Sign**: Is the relationship positive or negative? Check whether the coefficient has a negative value.
  - Positive coefficient: as the independent variable *increases*, the dependent variable is predicted to *increase*
  - Negative coefficient: as the independent variable *increases*, the dependent variable is predicted to *decrease*
  - Note about odds ratios: For certain types of (non-linear) regression, odds ratios (which always take on positive values) are sometimes displayed instead of coefficients; with an odds ratio, a value greater than one indicates a positive relationship while a value smaller than one indicates a negative relationship

- **Size**: How big is the (predictive) effect? This S is often the most difficult, and sometimes you may not have enough information to meaningfully evaluate it (e.g., if the units of measurement for a variable are not clearly explained).
  - For linear models: A one-unit increase in the independent variable predicts a β_i-unit change in the dependent variable (where β_i represents the value of the coefficent)
  - For non-linear models: Interpreting the size of a coefficient is typically more complicated than for a linear model; look for the authors' explanation of effect size or "magnitude" of association

Table 3.6 provides an example of regression results in a format similar to what you may encounter in many research publications. Note, however, that many publications do not list exact p-values; instead, they often use one or more asterisks (*) to denote coefficients with p-values smaller than 0.05 (sometimes also flagging p-values falling below various other relevant thresholds).

Table 3.6: Results for a regression with computer science GPA as the dependent variable.

	Coef.	Std. err.	p-value
verb_sat	0.0017	0.0010	0.10
math_sat	0.0048	0.0012	0.00014
(intercept)	-0.91	0.42	0.033
n	105		

	Coef.	Std. err.	p-value
r^2	0.487		

Up til now, we have only discussed simple linear regression, in which we have a single independent variable. But in Table 3.6, we find results for a regression where two independent variables—SAT scores on the verbal section (verb_sat) and SAT scores on the math section (math_sat)—are jointly used to predict students' GPA in computer science classes. It turns out that regression can easily be performed with multiple independent variables, as described in the appendix to this chapter. When we have multiple independent variables, we evaluate each one on its own terms when working through the three S's. For the results in Table 3.6, we can interpret the results as follows:

- verb_sat: The p-value for this variable (0.10) is greater than 0.05, so this variable is not statistically significant. Therefore, we don't necessarily need to interpret the sign or size. We might simply say that we could not establish a reliable link between verbal SAT scores and computer science GPA in this model.[10]
- math_sat: The p-value (0.00014) is smaller than 0.05, so math_sat is a statistically significant predictor of computer science GPA. The coefficient (0.0048) has a positive sign, so students with higher math SAT scores are predicted to have higher computer science GPAs. When it comes to size, a one-point increase in the math SAT (e.g., getting a 501 instead of a 500) predicts that the computer science GPA will be 0.0048 points higher. That seems very small, but a one-point increase on an SAT is barely noticeable (and not actually possible if scores are always multiples of ten). In this case, we can get a better sense of size if we consider an increase of 100 points in the math SAT, which requires multiplying the coefficient by 100. A 100-point increase in the math SAT (e.g., getting a 600 instead of a 500) predicts a computer science GPA that is 0.48 points higher. This is nearly half a grade point higher and would be quite noticeable to most students. Thus, the size of predictive effect now seems reasonably large.

Note that we do not need to apply the three S's to the intercept (sometimes labeled the "constant") because it is not a variable. Table 3.6 also contains some additional information frquently shown in regression tables: standard errors (which we will learn more about in Chapter 6), the sample size (n=105), and r-squared (a statistic often used to describe how well the regression model overall explains variation in the dependent variable).

Remember that the three S's are just a starting point. But they should be enough to help you follow along a little easier when reading the results sections of many research publications. If you've started working with a statistical software package by now, you can also try running your own models and seeing if you can use the three S's to help you understand the results.

[10]Note, however, that the absence of evidence is not necessarily evidence of absence. There could very well be a link between verbal SAT scores and computer science GPA—just one that we cannot reliably detect with this analysis (e.g., because our sample is too small to precisely estimate the association).

Chapter 3 Appendix: Multiple Regression[11]

In simple linear regression, a dependent variable is predicted from one independent variable. In multiple regression, the dependent variable is predicted by two or more variables. For example, in the SAT case study, you might want to predict a student's university grade point average on the basis of their High-School GPA (HSGPA) and their total SAT score (verbal + math). The basic idea is to find a linear combination[12] of HSGPA and SAT that best predicts University GPA (UGPA). That is, the problem is to find the values of b_1 and b_2 in the equation shown below that give the best predictions of UGPA. As in the case of simple linear regression, we define the best predictions as the predictions that minimize the squared errors of prediction.

$$\widehat{UGPA} = \alpha + \beta_1 HSGPA + \beta_2 SAT$$

where $\widehat{UGPA}$ is the predicted value of University GPA and α is a constant. For these data, the best prediction equation is shown below:

$$\widehat{UGPA} = 0.540 + 0.541 \times HSGPA + 0.008 \times SAT$$

In other words, to compute the prediction of a student's University GPA, you add up (a) 0.540, (b) their High-School GPA multiplied by 0.541, and (c) their SAT multiplied by 0.008. Table 3.7 shows the data and predictions for the first five students in the dataset.

Table 3.7: Data and Predictions

$HSGPA$	SAT	$\widehat{UGPA}$
3.45	1232	3.38
2.78	1070	2.89
2.52	1086	2.76

[11] This section is adapted from Rudy Guerra and David M. Lane. "Introduction to Multiple Regression." *Online Statistics Education: A Multimedia Course of Study.* https://onlinestatbook.com/2/regression/multiple_regression.html

[12] A linear combination of variables is a way of creating a new variable by combining other variables. A linear combination is one in which each variable is multiplied by a coefficient and the are products summed. For example, if

$$Y = 3X_1 + 2X_2 + .5X_3$$

then Y is a linear combination of the variables X_1, X_2, and X_3.

3.67	1287	3.55
3.24	1130	3.19

The values of β (β_1 and β_2) are called "regression coefficients."

The multiple correlation (R) is equal to the correlation between the predicted scores and the actual scores. In this example, it is the correlation between $\widehat{UGPA}$ and $UGPA$, which turns out to be 0.79. That is, R = 0.79. Note that R will never be negative since if there are negative correlations between the predictor variables and the criterion, the regression weights will be negative so that the correlation between the predicted and actual scores will be positive.

Interpretation of Regression Coefficients

A regression coefficient in multiple regression is the slope of the linear relationship between the criterion variable and the part of a predictor variable that is independent of all other predictor variables. In this example, the regression coefficient for HSGPA can be computed by first predicting HSGPA from SAT and saving the errors of prediction (the differences between $HSGPA$ and $\widehat{HSGPA}$). These errors of prediction are called "residuals" since they are what is left over in HSGPA after the predictions from SAT are subtracted, and represent the part of HSGPA that is independent of SAT. These residuals are referred to as HSGPA.SAT, which means they are the residuals in HSGPA after having been predicted by SAT. The correlation between HSGPA.SAT and SAT is necessarily 0.

The final step in computing the regression coefficient is to find the slope of the relationship between these residuals and UGPA. This slope is the regression coefficient for HSGPA. The following equation is used to predict HSGPA from SAT:

$$\widehat{HSGPA} = -1.314 + 0.0036 \times SAT$$

The residuals are then computed as:

$$HSGPA - \widehat{HSGPA}$$

The linear regression equation for the prediction of UGPA by the residuals is

$$\widehat{UGPA} = 3.173 + 0.541 \times HSGPA.SAT$$

Notice that the slope (0.541) is the same value given previously for β_1 in the multiple regression equation.

This means that the regression coefficient for HSGPA is the slope of the relationship between the dependent variable and the part of HSGPA that is independent of (uncorrelated with) the other independent variables. It represents the change in the dependent variable associated with a change of one in the independent variable when all other independent variables are held constant. Since the regression coefficient for HSGPA is 0.54, this means that, holding SAT constant, a change of one in HSGPA is associated with a change of 0.54 in $\widehat{UGPA}$. If two students had the same SAT and differed in HSGPA by 2, then you would predict they would differ in UGPA by (2)(0.54) = 1.08. Similarly, if they differed by 0.5, then you would predict they would differ by (0.50)(0.54) = 0.27.

The slope of the relationship between the dependent variable and the part of an independent variable that is unique from (independent of) other independent variables is its partial slope. Thus, the regression coefficient of 0.541 for HSGPA and the regression coefficient of 0.008 for SAT are partial slopes. Each partial slope represents the relationship between the independent variable and the dependent variable holding constant all of the other independent variables.

It is difficult to compare the coefficients for different variables directly because they are measured on different scales. A difference of 1 in HSGPA is a fairly large difference, whereas a difference of 1 on the SAT is negligible. Therefore, it can be advantageous to transform the variables so that they are on the same scale. The most straightforward approach is to standardize the variables (see Section 2.4.1) so that they each have a standard deviation of 1. A regression coefficient for standardized variables is called a "standardized coefficient" or "beta coefficient." For these data, the standardized coefficients are 0.625 and 0.198. These values represent the change in the dependent variable (in standard deviations) associated with a change of one standard deviation on an independent variable (holding constant the value(s) on the other independent variable(s)). Clearly, a change of one standard deviation on HSGPA is associated with a larger difference than a change of one standard deviation of SAT. In practical terms, this means that if you know a student's HSGPA, knowing the student's SAT does not aid the prediction of UGPA much. However, if you do not know the student's HSGPA, his or her SAT can aid in the prediction since the standardized coefficient in the simple regression predicting UGPA from SAT is 0.68. For comparison purposes, the standardized coefficient in the simple regression predicting UGPA from HSGPA is 0.78. As is typically the case, the partial slopes are smaller than the slopes in simple regression.

4 Estimation

4.1 Populations and Samples[1]

In statistics, we often rely on a **sample** — that is, a small subset of a larger set of data — to draw inferences about the larger set. The larger set is known as the **population** from which the sample is drawn.

> Example #1: You have been hired by the National Election Commission to examine how the American people feel about the fairness of the voting procedures in the U.S. Whom will you ask?

It is not practical to ask every single American how he or she feels about the fairness of the voting procedures. Instead, we query a relatively small number of Americans, and draw inferences about the entire country from their responses. The Americans actually queried constitute our sample of the larger population of all Americans. The mathematical procedures whereby we convert information about the sample into intelligent guesses about the population fall under the rubric of **inferential statistics**.

A sample is typically a small subset of the population. In the case of voting attitudes, we would sample a few thousand Americans drawn from the hundreds of millions that make up the country. In choosing a sample, it is therefore crucial that it not over-represent one kind of citizen at the expense of others. For example, something would be wrong with our sample if it happened to be made up entirely of Florida residents. If the sample held only Floridians, it could not be used to infer the attitudes of other Americans. The same problem would arise if the sample were comprised only of Republicans. Inferential statistics are based on the assumption that sampling is random. We trust a random sample to represent different segments of society in close to the appropriate proportions (provided the sample is large enough; see below).

> Example #2: We are interested in examining how many math classes have been taken on average by current graduating seniors at American colleges and universities during their four years in school. Whereas our population in the last example included all US citizens, now it involves just the graduating seniors throughout the

[1]This section is adapted from Mikki Hebl and David Lane. "Inferential Statistics." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/introduction/inferential.html

country. This is still a large set since there are thousands of colleges and universities, each enrolling many students. (New York University, for example, enrolls 48,000 students.) It would be prohibitively costly to examine the transcript of every college senior. We therefore take a sample of college seniors and then make inferences to the entire population based on what we find. To make the sample, we might first choose some public and private colleges and universities across the United States. Then we might sample 50 students from each of these institutions. Suppose that the average number of math classes taken by the people in our sample were 3.2. Then we might speculate that 3.2 approximates the number we would find if we had the resources to examine every senior in the entire population. But we must be careful about the possibility that our sample is non-representative of the population. Perhaps we chose an overabundance of math majors, or chose too many technical institutions that have heavy math requirements. Such bad sampling makes our sample unrepresentative of the population of all seniors.

Example #3: A substitute teacher wants to know how students in the class did on their last test. The teacher asks the 10 students sitting in the front row to state their latest test score. He concludes from their report that the class did extremely well. What is the sample? What is the population? Can you identify any problems with choosing the sample in the way that the teacher did?

In Example #3, the population consists of all students in the class. The sample is made up of just the 10 students sitting in the front row. The sample is not likely to be representative of the population. Those who sit in the front row tend to be more interested in the class and tend to perform higher on tests. Hence, the sample may perform at a higher level than the population.

Example #4: A coach is interested in how many cartwheels the average college freshmen at his university can do. Eight volunteers from the freshman class step forward. After observing their performance, the coach concludes that college freshmen can do an average of 16 cartwheels in a row without stopping.

In Example #4, the population is the class of all freshmen at the coach's university. The sample is composed of the 8 volunteers. The sample is poorly chosen because volunteers are more likely to be able to do cartwheels than the average freshman; people who can't do cartwheels probably did not volunteer! In the example, we are also not told of the gender of the volunteers. Were they all women, for example? That might affect the outcome, contributing to the non-representative nature of the sample (if the school is co-ed).

4.1.1 Simple Random Sampling

Researchers adopt a variety of sampling strategies. The most straightforward is **simple random sampling**. Such sampling requires every member of the population to have an equal

chance of being selected into the sample. In addition, the selection of one member must be independent of the selection of every other member. That is, picking one member from the population must not increase or decrease the probability of picking any other member (relative to the others). In this sense, we can say that simple random sampling chooses a sample by pure chance. To check your understanding of simple random sampling, consider the following example. What is the population? What is the sample? Was the sample picked by simple random sampling? Is it biased?

> Example #5: A research scientist is interested in studying the experiences of twins raised together versus those raised apart. She obtains a list of twins from the National Twin Registry, and selects two subsets of individuals for her study. First, she chooses all those in the registry whose last name begins with Z. Then she turns to all those whose last name begins with B. Because there are so many names that start with B, however, our researcher decides to incorporate only every other name into her sample. Finally, she mails out a survey and compares characteristics of twins raised apart versus together.

In Example #5, the population consists of all twins recorded in the National Twin Registry. It is important that the researcher only make statistical generalizations to the twins on this list, not to all twins in the nation or world. That is, the National Twin Registry may not be representative of all twins. Even if inferences are limited to the Registry, a number of problems affect the sampling procedure we described. For instance, choosing only twins whose last names begin with Z does not give every individual an equal chance of being selected into the sample. Moreover, such a procedure risks over-representing ethnic groups with many surnames that begin with Z. There are other reasons why choosing just the Z's may bias the sample. Perhaps such people are more patient than average because they often find themselves at the end of the line! The same problem occurs with choosing twins whose last name begins with B. An additional problem for the B's is that the "every-other-one" procedure disallowed adjacent names on the B part of the list from being both selected. Just this defect alone means the sample was not formed through simple random sampling.

4.1.2 Sample size matters

Recall that the definition of a random sample is a sample in which every member of the population has an equal chance of being selected. This means that the sampling procedure rather than the results of the procedure define what it means for a sample to be random. Random samples, especially if the sample size is small, are not necessarily representative of the entire population. For example, if a random sample of 20 subjects were taken from a population with an equal number of males and females, there would be a nontrivial probability (0.06) that 70% or more of the sample would be female. Such a sample would not be representative, although it would be drawn randomly. Only a large sample size makes it likely that our sample is close to representative of the population. For this reason, inferential statistics take

into account the sample size when generalizing results from samples to populations. In later chapters, you'll see what kinds of mathematical techniques ensure this sensitivity to sample size.

4.1.3 More complex sampling

Sometimes it is not feasible to build a sample using simple random sampling. To see the problem, consider the fact that both Dallas and Houston are competing to be hosts of the 2012 Olympics. Imagine that you are hired to assess whether most Texans prefer Houston to Dallas as the host, or the reverse. Given the impracticality of obtaining the opinion of every single Texan, you must construct a sample of the Texas population. But now notice how difficult it would be to proceed by simple random sampling. For example, how will you contact those individuals who don't vote and don't have a phone? Even among people you find in the telephone book, how can you identify those who have just relocated to California (and had no reason to inform you of their move)? What do you do about the fact that since the beginning of the study, an additional 4,212 people took up residence in the state of Texas? As you can see, it is sometimes very difficult to develop a truly random procedure. For this reason, other kinds of sampling techniques have been devised. We now discuss two of them.

4.1.4 Random Assignment

In experimental research, populations are often hypothetical. For example, in an experiment comparing the effectiveness of a new anti-depressant drug with a placebo, there is no actual population of individuals taking the drug. In this case, a specified population of people with some degree of depression is defined and a random sample is taken from this population. The sample is then randomly divided into two groups; one group is assigned to the treatment condition (drug) and the other group is assigned to the control condition (placebo). This random division of the sample into two groups is called **random assignment**. Random assignment is critical for the validity of an experiment. For example, consider the bias that could be introduced if the first 20 subjects to show up at the experiment were assigned to the experimental group and the second 20 subjects were assigned to the control group. It is possible that subjects who show up late tend to be more depressed than those who show up early, thus making the experimental group less depressed than the control group even before the treatment was administered.

In experimental research of this kind, failure to assign subjects randomly to groups is generally more serious than having a non-random sample. Failure to randomize (the former error) invalidates the experimental findings. A non-random sample (the latter error) simply restricts the generalizability of the results.

4.1.5 Stratified Sampling

Since simple random sampling often does not ensure a representative sample, a sampling method called **stratified random sampling** is sometimes used to make the sample more representative of the population. This method can be used if the population has a number of distinct "strata" or groups. In stratified sampling, you first identify members of your sample who belong to each group. Then you randomly sample from each of those subgroups in such a way that the sizes of the subgroups in the sample are proportional to their sizes in the population.

Let's take an example: Suppose you were interested in views of capital punishment at an urban university. You have the time and resources to interview 200 students. The student body is diverse with respect to age; many older people work during the day and enroll in night courses (average age is 39), while younger students generally enroll in day classes (average age of 19). It is possible that night students have different views about capital punishment than day students. If 70% of the students were day students, it makes sense to ensure that 70% of the sample consisted of day students. Thus, your sample of 200 students would consist of 140 day students and 60 night students. The proportion of day students in the sample and in the population (the entire university) would be the same. Inferences to the entire population of students at the university would therefore be more secure.

4.2 Confidence Intervals[2]

Say you were interested in the mean weight of 10-year-old girls living in the United States. Since it would have been impractical to weigh all the 10-year-old girls in the United States, you took a sample of 16 and found that the mean weight was 90 pounds. This sample mean of 90 is a **point estimate** of the population mean. A point estimate by itself is of limited usefulness because it does not reveal the uncertainty associated with the estimate; you do not have a good sense of how far this sample mean may be from the population mean. For example, can you be confident that the population mean is within 5 pounds of 90? You simply do not know.

Confidence intervals provide more information than point estimates. Confidence intervals for means are intervals constructed using a procedure that will contain the population mean a specified proportion of the time, typically either 95% or 99% of the time. These intervals are referred to as 95% and 99% confidence intervals respectively. An example of a 95% confidence interval is shown below:

$$72.85 < \mu < 107.15$$

[2]This section is adapted from David M. Lane. "Confidence Intervals Introduction." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/estimation/confidence.html

There is good reason to believe that the population mean lies between these two bounds of 72.85 and 107.15 since 95% of the time confidence intervals contain the true mean.

If repeated samples were taken and the 95% confidence interval computed for each sample, 95% of the intervals would contain the population mean. Naturally, 5% of the intervals would not contain the population mean.

It is natural to interpret a 95% confidence interval as an interval with a 0.95 probability of containing the population mean. However, the proper interpretation is not that simple (as will be explained more fully in Section 4.4). One problem is that the computation of a confidence interval does not take into account any other information you might have about the value of the population mean. For example, if numerous prior studies had all found sample means above 110, it would not make sense to conclude that there is a 0.95 probability that the population mean is between 72.85 and 107.15.

Confidence intervals can be computed for various **parameters**,[3] not just the mean. For example, it is common to compute a confidence interval for ρ, the population value of Pearson's r, based on sample data.

4.3 Using Confidence Intervals[4]

It is much more common for a researcher to be interested in the difference between means than in the specific values of the means themselves. Using statistical jargon introduced in the prior section, we could therefore say that the *parameter of interest* is often the difference in population means. We take as an example the data from the "Animal Research"[5] case study. In this experiment, students rated (on a 7-point scale) whether they thought animal research is wrong. The sample sizes, means, and variances are shown separately for males and females in Table 4.1.

Table 4.1: Means and Variances in Animal Research study.

Condition	n	Mean	Variance
Females	17	5.353	2.743
Males	17	3.882	2.985

As you can see, the females rated animal research as more wrong than did the males. This sample difference between the female mean of 5.35 and the male mean of 3.88 is 1.47. However,

[3] A parameter is a value calculated in a population. For example, the mean of the numbers in a population is a parameter. Compare with a sample statistic, which is a value computed in a sample to estimate a parameter. (https://onlinestatbook.com/2/glossary/parameter.html)

[4] This section is adapted from David M. Lane. "Difference between Means." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/estimation/difference_means.html

[5] http://onlinestatbook.com/2/case_studies/animal_research.html

the gender difference in this particular sample is not very important. What is important is the difference in the population. The difference in sample means is used to estimate the difference in population means. The accuracy of the estimate is revealed by a confidence interval.

In order to construct a confidence interval, we are going to make some assumptions. These won't make a lot of sense yet, but here are the three assumptions we need to make:

1. The two populations have the same variance. This assumption is called the assumption of homogeneity of variance.
2. The populations are normally distributed.
3. Each value is sampled independently from each other value.

Using these assumptions, one can use a bunch of fancy math formulas (or statistical software) to get the following confidence interval:

$$0.29 \leq \mu_f - \mu_m \leq 2.65$$

where μ_f is the population mean for females and μ_m is the population mean for males. Since the difference in these population means is the main parameter we wish to estimate, we could say the confidence interval for the parameter of interest is [0.29, 2.65]. Because all values within this range are positive, this analysis provides evidence that the mean for females is higher than the mean for males. More specifically, the difference between means in the population is likely to be between 0.29 and 2.65. Note that since 0 does not fall within the range of the confidence interval, this suggests that there *is* a difference between males and females, so we can also say that the results are **statistically significant**.

If, instead, we had found a confidence interval of [-1.03, 2.65], we could not rule out the possibility of no difference between males and females, since 0 falls between -1.03 and 2.65.

4.4 Interpreting Confidence Intervals Correctly[6]

While confidence intervals are very useful, it is difficult to provide a technically-accurate interpretation of one.[7] That is because the "% confidence" in a "95% confidence interval for the mean" refers to the accuracy of the *process* of creating a confidence interval—not the probability that a specific confidence interval we encounter will contain the true value of the population mean. If this distinction seems confusing, it is!

[6]This section is written by Nathan Favero.

[7]Stephens, M. (2023). The Bayesian lens and Bayesian blinkers. *Philosophical Transactions of the Royal Society A*, *381*(2247), 20220144.

Kass, R. E. (2011). Statistical inference: The big picture. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *26*(1), 1.

Fortunately, even if you miss the precise details, you will still probably get something useful out of confidence intervals.[8] Nonetheless, let's try to set the record straight.

An analogy may help. Suppose you are interacting with a chatbot that is truthful 95% of the time and lies the other 5%.[9] For each statement, will you always conclude it has a 95% chance of being true? Not necessarily. If the chatbot discusses a topic you already know a lot about, you will probably be able to pick out the lies from the true statements with fairly high confidence. Some things the bot says will be things you know to be true, so you can be nearly 100% sure they are true. Other statements will be things you're quite sure are wrong, so you will conclude that the probability they are true is close to 0%. If you wanted to be very systematic, you could even use the mathematical formula known as Bayes' theorem[10] to combine your prior knowledge of a statement's probability of being true with the fact that a 95%-accurate bot claimed the statement was true, allowing you to precisely quantify how confident you should be about the statement's truth in the end.

Now imagine you ask this same bot to start telling you about a topic you know nothing about. Absent any prior insights into which statements are likely to be true or false, it would now be reasonable to conclude that each statement the bot makes has a 95% chance of being true.

In the same way, it turns out that *absent any other information,* a 95% confidence interval is often a good approximation for a range of values that contains the population parameter with 95% probability.[11] Thus, I think it is quite reasonable that many of us, when we see a mean estimate with a 95% confidence interval ranging from A to B, assume there is a 95% chance the population mean does indeed lie between A and B. But technically, that is not a direct interpretation of the confidence interval; instead, this statement about plausible values of the population mean is a subjective conclusion that I can draw based on the confidence interval. Another person might see the same confidence interval and reasonably decide—drawing on their own prior knowledge of the topic—that the confidence interval contains values that are highly implausible, and thus they would reach a different conclusion from me about how likely the interval is to contain the true population mean.

If you want to be precise in how you interpret a confidence interval, you can make the following statement any time you encounter a confidence interval with a range of [A, B]:

---

[8] Anderson, A. A. (2019). Assessing statistical results: magnitude, precision, and model uncertainty. *The American Statistician*, *73*(sup1), 118-121.

[9] This example is adapted from Behar, R., Grima, P., & Marco-Almagro, L. (2013). Twenty-five analogies for explaining statistical concepts. *The American Statistician*, *67*(1), 44-48.

[10] See https://onlinestatbook.com/2/glossary/bayes.html or https://onlinestatbook.com/2/probability/bayes_demo.html.

[11] Kass, R. E. (2011). Statistical inference: The big picture. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *26*(1), 1.

Albers, C. J., Kiers, H. A., & van Ravenzwaaij, D. (2018). Credible confidence: A pragmatic view on the frequentist vs Bayesian debate. *Collabra: Psychology*, *4*(1), 31.

Greenland, S., & Poole, C. (2013). Living with p values: resurrecting a Bayesian perspective on frequentist statistics. *Epidemiology*, *24*(1), 62-68.

> Using a process with 95% accuracy (in theory), it is estimated that the parameter lies between A and B.

If you want to elaborate on how the confidence interval can inform our practical understanding, you might add that:

> Assuming no additional information and an appropriate statistical model, this result usually suggests that we can be about 95% confident the parameter lies between A and B.

5 Probability Distributions

5.1 Various Types of Distributions[1]

5.1.1 Distributions of Discrete Variables

I recently purchased a bag of Plain M&M's. The M&M's were in six different colors. A quick count showed that there were 55 M&M's: 17 brown, 18 red, 7 yellow, 7 green, 2 blue, and 4 orange. These counts are shown below in Table 5.1.

Table 5.1: Frequencies in the Bag of M&M's

Color	**Frequency**
Brown	17
Red	18
Yellow	7
Green	7
Blue	2
Orange	4

This table is called a frequency table and it describes the distribution of M&M color frequencies. Not surprisingly, this kind of distribution is called a **frequency distribution**. Often a frequency distribution is shown graphically as in Figure 5.1.

The distribution shown in Figure 5.1 concerns just my one bag of M&M's. You might be wondering about the distribution of colors for all M&M's. The manufacturer of M&M's provides some information about this matter, but they do not tell us exactly how many M&M's of each color they have ever produced. Instead, they report proportions rather than frequencies. Figure 5.2 shows these proportions. Since every M&M is one of the six familiar colors, the six proportions shown in the figure add to one. We call Figure 5.2 a **probability distribution** because if you choose an M&M at random, the probability of getting, say, a brown M&M is equal to the proportion of M&M's that are brown (0.30).

Notice that the distributions in Figure 5.1 and Figure 5.2 are not identical. Figure 5.1 portrays the distribution in a sample of 55 M&M's. Figure 5.2 shows the proportions for all M&M's. Chance factors involving the machines used by the manufacturer introduce random variation

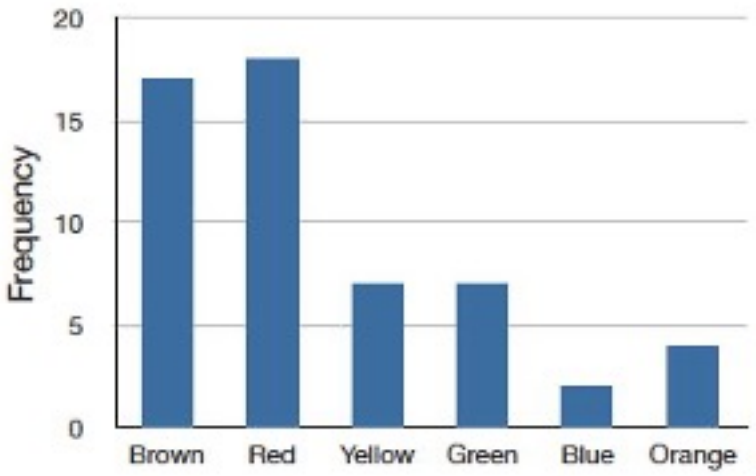


Figure 5.1: Distribution of 55 M&M's.

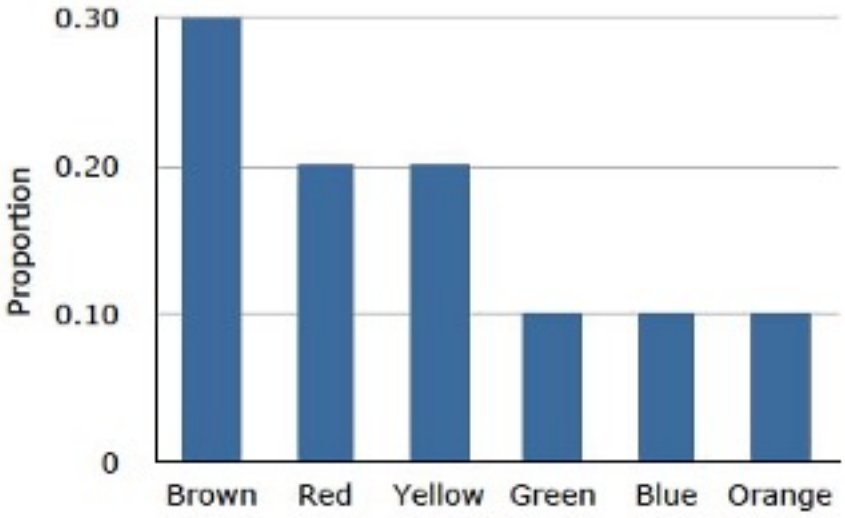


Figure 5.2: Distribution of all M&M's.

into the different bags produced. Some bags will have a distribution of colors that is close to Figure 5.2; others will be further away.

5.1.2 Continuous Variables

The variable "color of M&M" used in this example is a discrete variable, and its distribution is also called discrete. Let us now extend the concept of a distribution to continuous variables.

The data shown in Table 5.2 are the times it took David Lane (the author of much of the material appearing in this book) to move the cursor over a small target in a series of 20 trials. The times are sorted from shortest to longest. The variable "time to respond" is a continuous variable. With time measured accurately (to many decimal places), no two response times would be expected to be the same. Measuring time in milliseconds (thousandths of a second) is often precise enough to approximate a continuous variable in psychology. As you can see in Table 5.2, measuring David Lane's responses this way produced times no two of which were the same. As a result, a frequency distribution would be uninformative: it would consist of the 20 times in the experiment, each with a frequency of 1.

Table 5.2: Response Times

568	720
577	728
581	729
640	777
641	808
645	824
657	825
673	865
696	875
703	1007

The solution to this problem is to create a grouped frequency distribution, as we saw when learning about histograms in Chapter 1. In a grouped frequency distribution, scores falling within various ranges are tabulated. Table 5.3 shows a grouped frequency distribution for these 20 times.

Table 5.3: Grouped frequency distribution

Range	**Frequency**

500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

Figure 5.3 shows a histogram for the frequency distribution in Table 5.3.

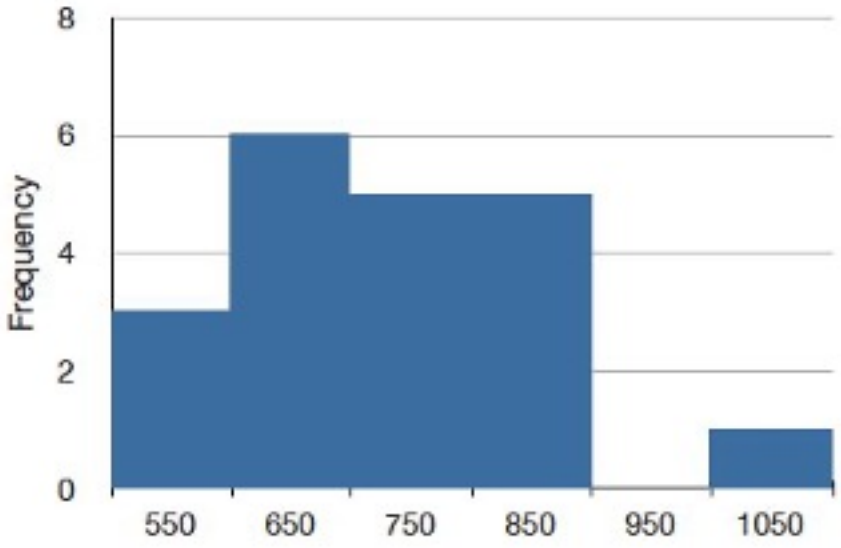


Figure 5.3: A histogram of the grouped frequency distribution shown in Table 5.3. The labels on the X-axis are the middle values of the range they represent.

5.1.3 Probability Densities

The histogram in Figure 5.3 portrays just David Lane's 20 times in the one experiment. To represent the probability associated with an arbitrary movement (which can take any positive amount of time), we must represent all these potential times at once. For this purpose, we plot the distribution for the continuous variable of time. Distributions for continuous variables are called continuous distributions. They also carry the fancier name **probability density**. Some probability densities have particular importance in statistics. A very important one is shaped like a bell, and called the **normal distribution**. Many naturally-occurring phenomena can be approximated surprisingly well by this distribution. It will serve to illustrate some features of all continuous distributions.

An example of a normal distribution is shown in Figure 5.4. Do you see the "bell"? The normal distribution doesn't represent a real bell, however, since the left and right tips extend indefinitely (we can't draw them any further so they look like they've stopped in our diagram). The Y-axis in the normal distribution represents the "density of probability." Intuitively, it shows the chance of obtaining values near corresponding points on the X-axis. In Figure 5.4, for example, the probability of an observation with value near 40 is about half of the probability of an observation with value near 50.

Although this text does not discuss the concept of probability density in detail, you should keep the following ideas in mind about the curve that describes a continuous distribution (like the normal distribution). First, the area under the curve equals 1. Second, the probability of any exact value of X is 0. Finally, the area under the curve and bounded between two given points on the X-axis is the probability that a number chosen at random will fall between the two points. Let us illustrate with David Lane's hand movements. First, the probability that his movement takes some amount of time is one! (We exclude the possibility of him never finishing his gesture.) Second, the probability that his movement takes exactly 598.956432342346576 milliseconds is essentially zero. (We can make the probability as close as we like to zero by making the time measurement more and more precise.) Finally, suppose that the probability of David Lane's movement taking between 600 and 700 milliseconds is one tenth. Then the continuous distribution for David Lane's possible times would have a shape that places 10% of the area below the curve in the region bounded by 600 and 700 on the X-axis.

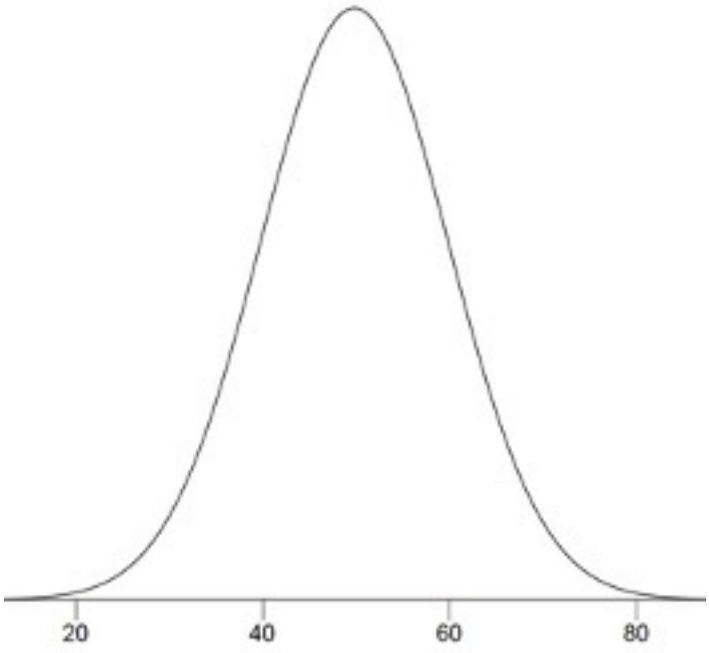


Figure 5.4: A normal distribution.

5.1.4 Shapes of Distributions

As we've already seen when graphing different data, distributions have different shapes; they don't all look like the normal distribution in Figure 5.4. For example, the normal probability density is higher in the middle compared to its two tails. Other distributions need not have this feature. There is even variation among the distributions that we call "normal." For example, some normal distributions are more spread out than the one shown in Figure 5.4 (their tails begin to hit the X-axis further from the middle of the curve – for example, at 10 and 90 if drawn in place of Figure 5.4). Others are less spread out (their tails might approach the X-axis at 30 and 70). We'll learn more about the details of the normal distribution later in this chapter.

The distribution shown in Figure 5.4 is symmetric; if you folded it in the middle, the two sides would match perfectly. Figure 5.5 shows the discrete distribution of scores on a psychology test. This distribution is not symmetric: the tail in the positive direction extends further than the tail in the negative direction. A distribution with the longer tail extending in the positive direction is said to have a **positive skew**. It is also described as "skewed to the right."

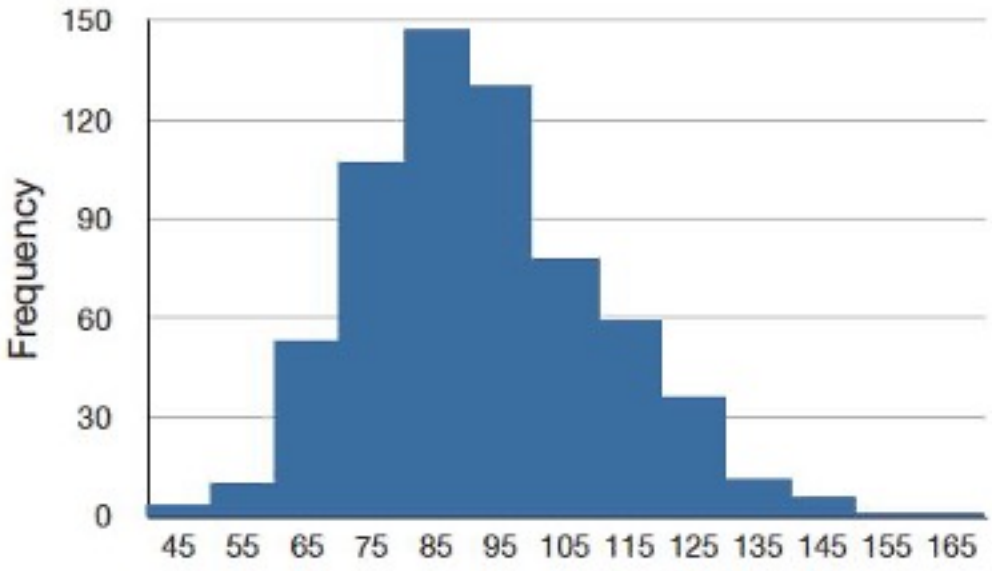


Figure 5.5: A distribution with a positive skew.

Figure 5.6 shows the salaries of major league baseball players in 1974 (in thousands of dollars). This distribution has an extreme positive skew.

A continuous distribution with a positive skew is shown in Figure 5.7.

Although less common, some distributions have a **negative skew**. Figure 5.8 shows the scores on a 20-point problem on a statistics exam. Since the tail of the distribution extends to the

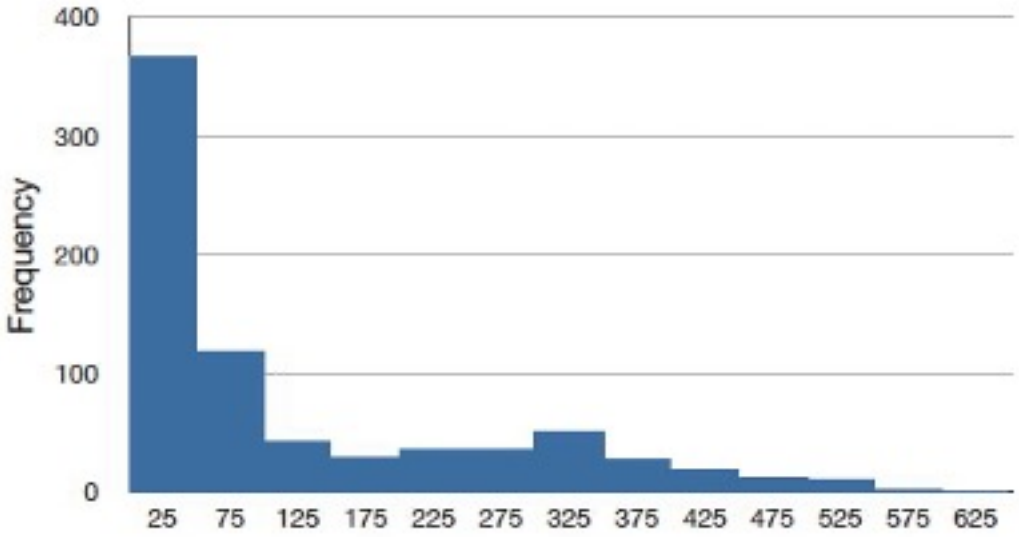


Figure 5.6: A distribution with a very large positive skew.

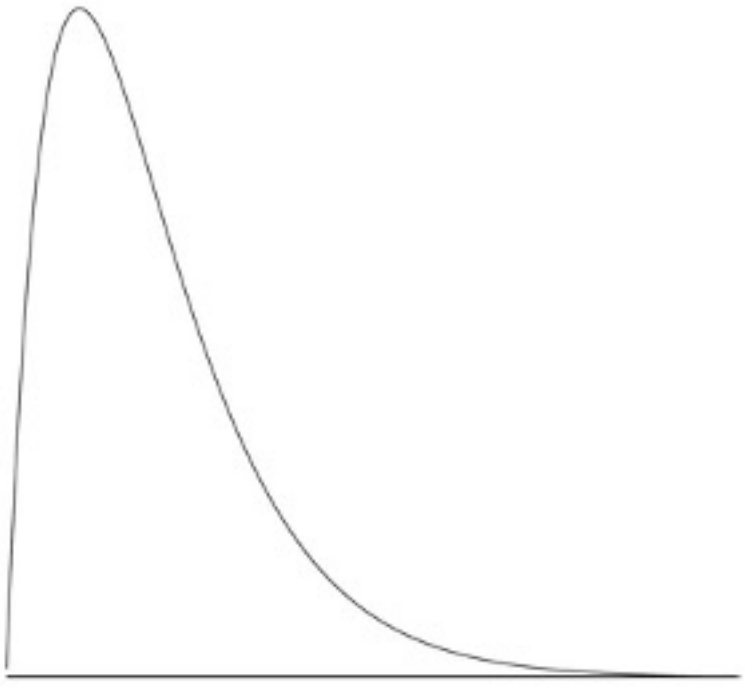


Figure 5.7: A continuous distribution with a positive skew.

left, this distribution is skewed to the left.

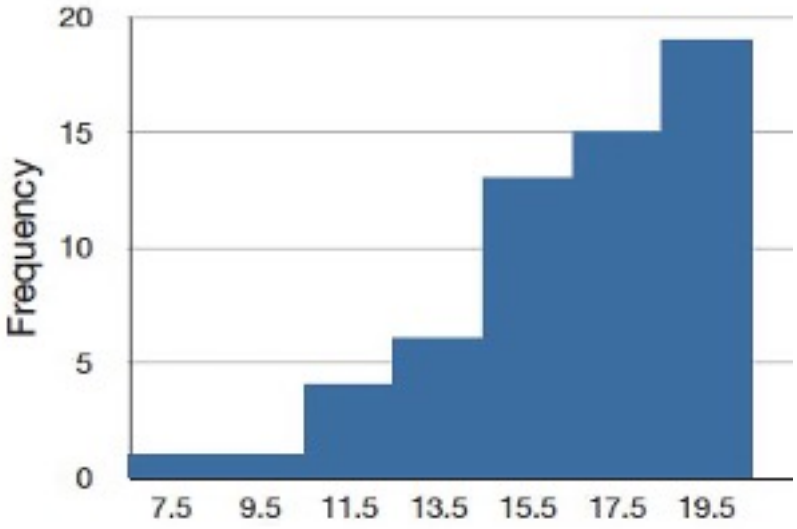


Figure 5.8: A distribution with negative skew.

A continuous distribution with a negative skew is shown in Figure 5.9.

The distributions shown so far all have one distinct high point or peak. The distribution in Figure 5.10 has two distinct peaks. A distribution with two peaks is called a **bimodal distribution**.

5.2 Normal Distributions[2]

The normal distribution is the most important and most widely used distribution in statistics. It is sometimes called the "bell curve," although the tonal qualities of such a bell would be less than pleasing. It is also called the "Gaussian curve" after the mathematician Karl Friedrich Gauss. Although Gauss played an important role in its history, Abraham de Moivre first discovered the normal distribution.

Strictly speaking, it is not correct to talk about "the normal distribution" since there are many normal distributions. Normal distributions can differ in their means and in their standard deviations. Figure 5.4 shows three normal distributions. The green (left-most) distribution has a mean of -3 and a standard deviation of 0.5, the distribution in red (the middle distribution) has a mean of 0 and a standard deviation of 1, and the distribution in black (right-most) has a mean of 2 and a standard deviation of 3. These as well as all other normal distributions are symmetric with relatively more values at the center of the distribution and relatively few in the tails.

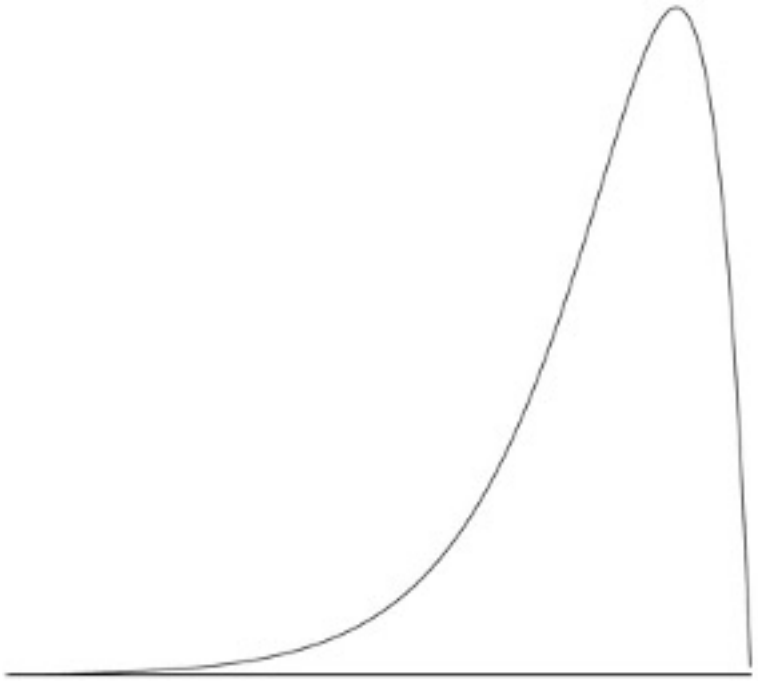


Figure 5.9: A continuous distribution with a negative skew.

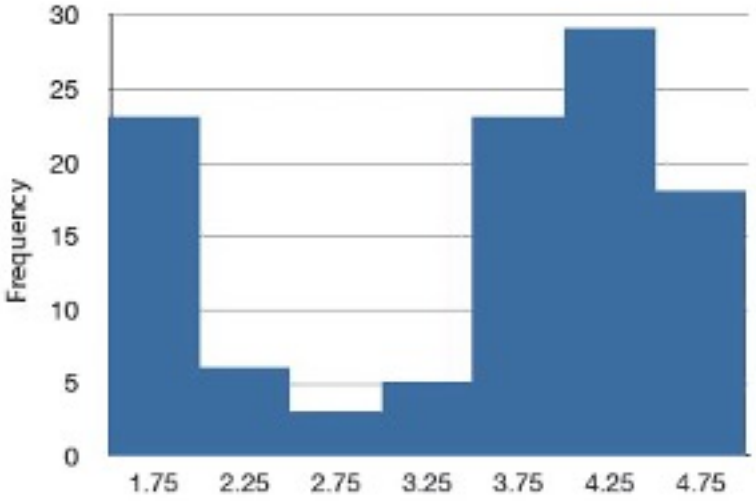


Figure 5.10: Frequencies of times between eruptions of the Old Faithful geyser. Notice the two distinct peaks: one at 1.75 and the other at 4.25.

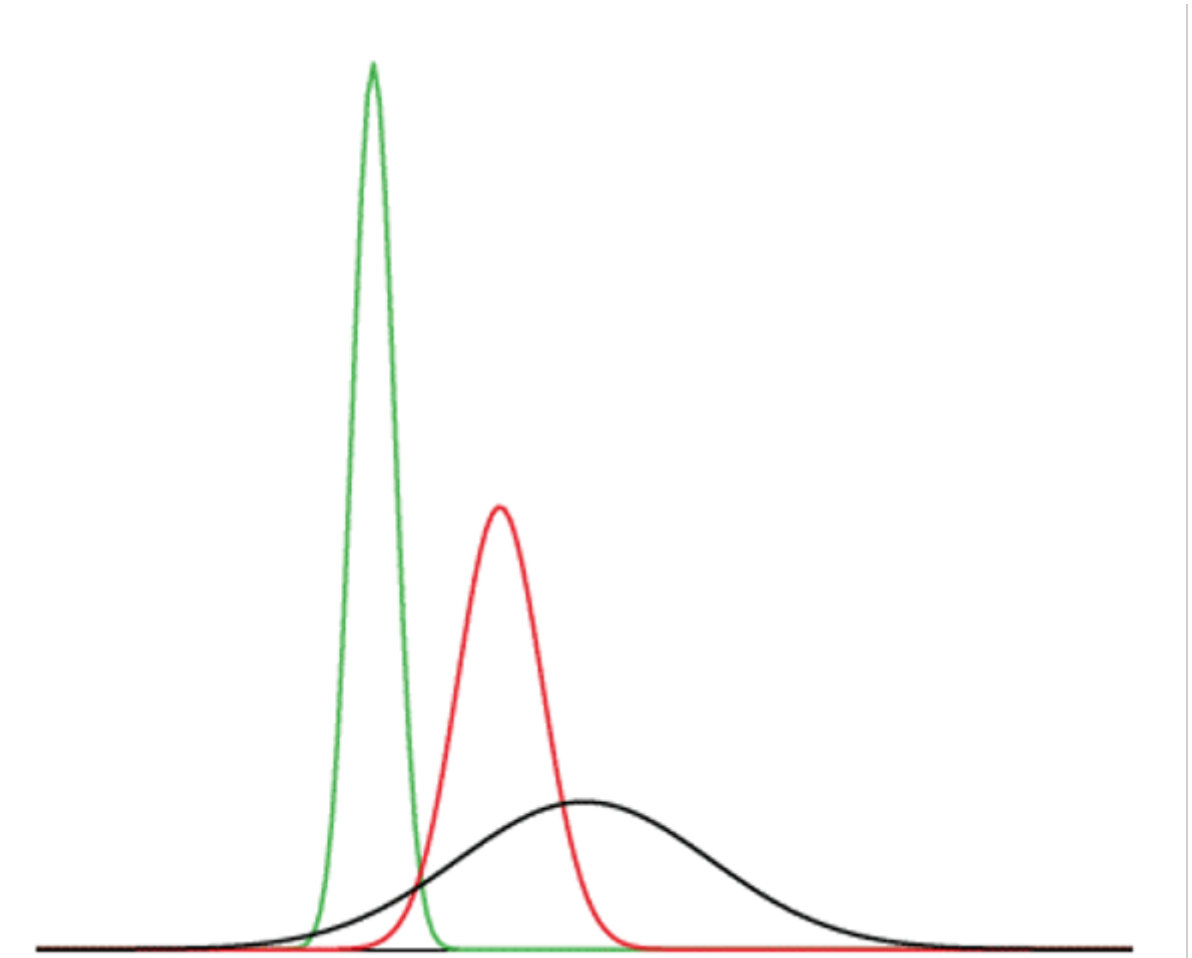


Figure 5.11: Normal distributions differing in mean and standard deviation.

Seven features of normal distributions are listed below. These features are illustrated in more detail in the remaining sections of this chapter.

1. Normal distributions are symmetric around their mean.
2. The mean, median, and mode of a normal distribution are equal.
3. The area under the normal curve is equal to 1.0.
4. Normal distributions are denser in the center and less dense in the tails.
5. Normal distributions are defined by two parameters, the mean ( ) and the standard deviation ( ).
6. 68% of the area of a normal distribution is within one standard deviation of the mean.

7. Approximately 95% of the area of a normal distribution is within two standard deviations of the mean.

5.2.1 Importance of Normal Distributions[3]

The importance of the normal curve stems primarily from the fact that the distributions of many natural phenomena are at least approximately normally distributed. One of the first applications of the normal distribution was to the analysis of errors of measurement made in astronomical observations, errors that occurred because of imperfect instruments and imperfect observers. Galileo in the 17th century noted that these errors were symmetric and that small errors occurred more frequently than large errors. This led to several hypothesized distributions of errors, but it was not until the early 19th century that it was discovered that these errors followed a normal distribution. Independently, the mathematicians Adrain in 1808 and Gauss in 1809 developed the formula for the normal distribution and showed that errors were fit well by this distribution.

Most statistical procedures for testing differences between means assume normal distributions. Because the distribution of means is very close to normal, these tests work well even if the original distribution is only roughly normal.

Quételet was the first to apply the normal distribution to human characteristics. He noted that characteristics such as height, weight, and strength were normally distributed.

5.2.2 Areas Under Normal Distributions[4]

Areas under portions of a normal distribution can be computed by using calculus. Since this is a non-mathematical treatment of statistics, we will rely on computer programs and tables to determine these areas. Figure 5.12 shows a normal distribution with a mean of 50 and a standard deviation of 10. The shaded area between 40 and 60 contains 68% of the distribution.

Figure 5.13 shows a normal distribution with a mean of 100 and a standard deviation of 20. Figure 5.1, 68% of the distribution is within one standard deviation of the mean.

The normal distributions shown in Figure 5.12 and Figure 5.13 are specific examples of the general rule that ***68% of the area of any normal distribution is within one standard deviation of the mean***.

Figure 5.14 shows a normal distribution with a mean of 75 and a standard deviation of 10. The shaded area contains 95% of the area and extends from 55.4 to 94.6. ***For all normal distributions, 95% of the area is within 1.96 standard deviations of the mean***. For quick approximations, it is sometimes useful to round off and use 2 rather than 1.96 as the number of standard deviations you need to extend from the mean so as to include 95% of the area.

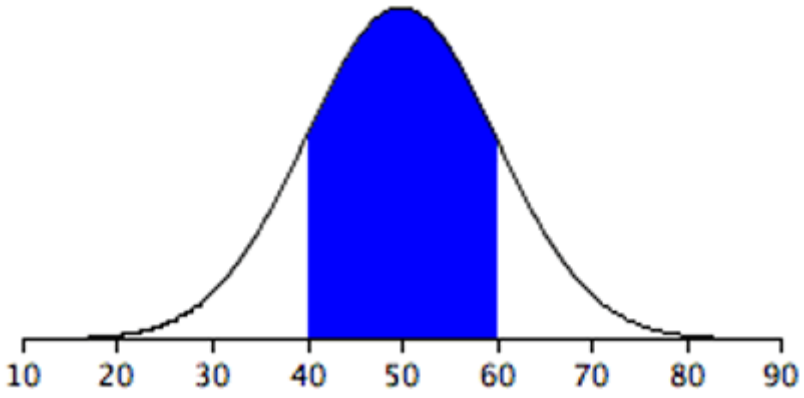


Figure 5.12: Normal distribution with a mean of 50 and standard deviation of 10. 68% of the area is within one standard deviation (10) of the mean (50).

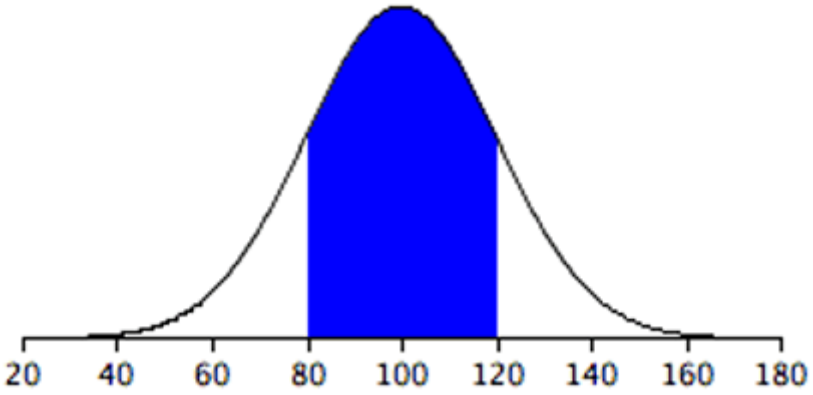


Figure 5.13: Normal distribution with a mean of 100 and standard deviation of 20. 68% of the area is within one standard deviation (20) of the mean (100).

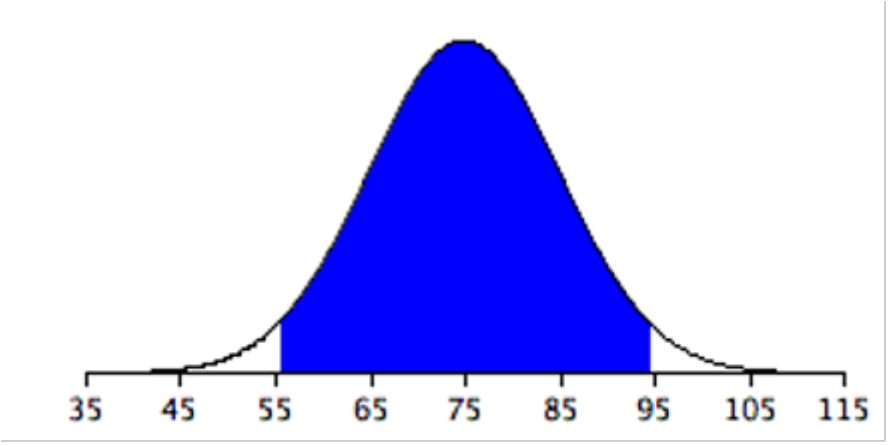


Figure 5.14: A normal distribution with a mean of 75 and a standard deviation of 10. 95% of the area is within 1.96 standard deviations of the mean.

It is easy to find free online normal distribution calculators that will give you the areas under the normal distribution (e.g., http://onlinestatbook.com/2/calculators/normal_dist.html). For example, you can use one to find the proportion of a normal distribution with a mean of 90 and a standard deviation of 12 that is above 110. Set the mean to 90 and the standard deviation to 12. Then enter "110" in the box to the right of the radio button "Above." At the bottom of the display you will see that the shaded area is 0.0478. See if you can use the calculator to find that the area between 115 and 120 is 0.0124.

Say you wanted to find the score corresponding to the 75th percentile of a normal distribution with a mean of 90 and a standard deviation of 12. Using an inverse normal calculator (e.g., http://onlinestatbook.com/2/calculators/inverse_normal_dist.html), you enter the parameters as shown in Figure 5.16 and find that the area below 98.09 is 0.75.

5.2.3 The Standard Normal Distribution[5]

As discussed above, normal distributions do not necessarily have the same means and standard deviations. A normal distribution with a mean of 0 and a standard deviation of 1 is called a ***standard normal distribution***.

---

[1] This section is adapted from David M. Lane and Heidi Ziemer. "Distributions." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/introduction/distributions.html

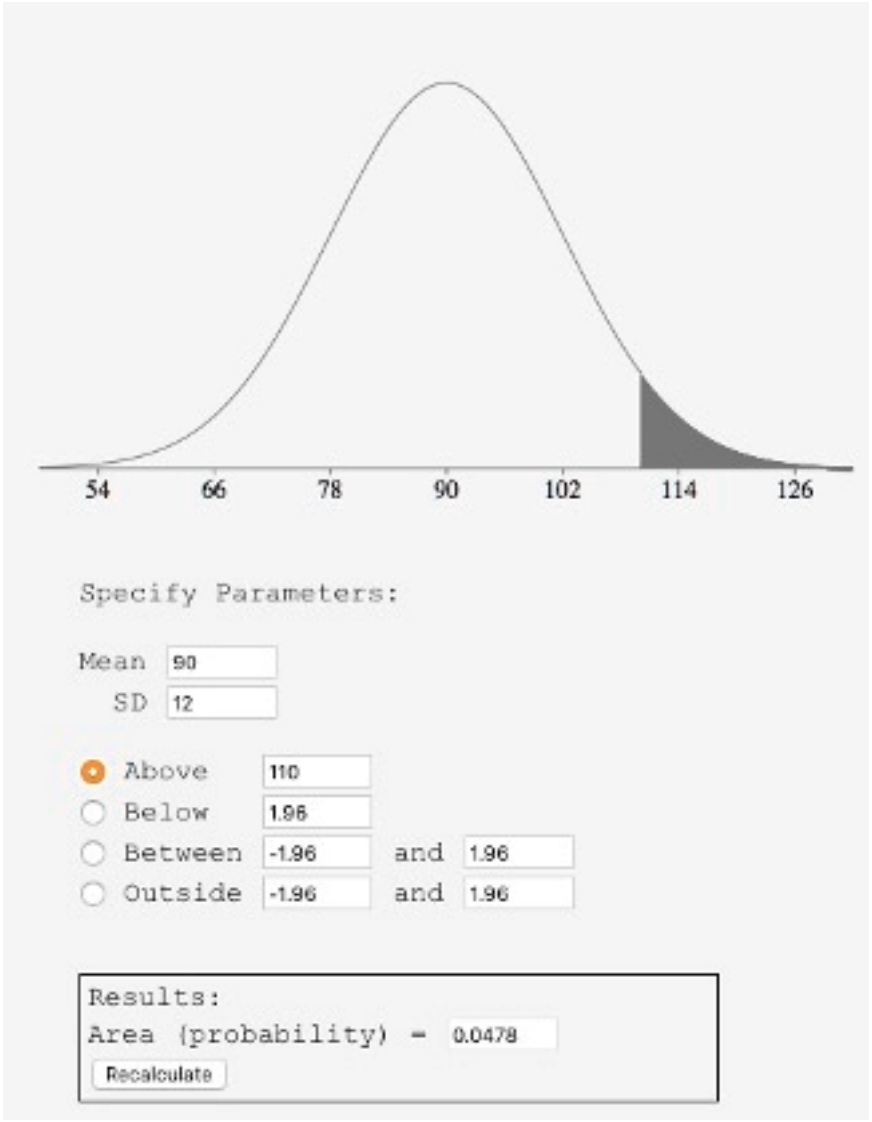


Figure 5.15: Display from calculator showing the area above 110.

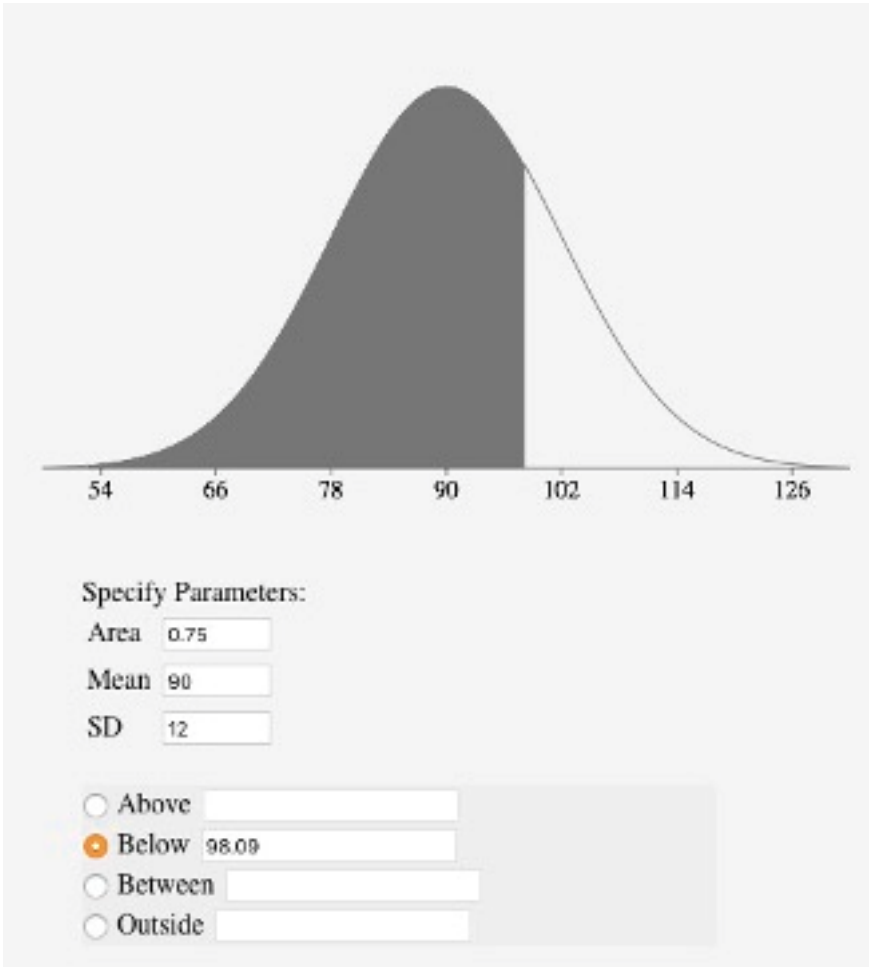


Figure 5.16: Display from normal calculator showing that the 75th percentile is 98.09.

[2] The initial part of this section is adapted from David M. Lane. "Introduction to Normal Distributions." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/normal_distribution/intro.html

[3] This subsection is adapted from David M. Lane. "History of the Normal Distribution." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/normal_distribution/history_normal.html

[4] This subsection is adapted from David M. Lane. "Areas Under Normal Distributions." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/normal_distribution/areas_normal.html

[5] This subsection is adapted from David M. Lane. "Standard Normal Distribution." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/normal_distribution/standard_normal.html

6 Sampling Distributions

6.1 Introduction to Sampling Distributions[1]

Suppose you randomly sampled 10 people from the population of women in Houston, Texas, between the ages of 21 and 35 years and computed the mean height of your sample. You would not expect your sample mean to be equal to the mean of all women in Houston. It might be somewhat lower or it might be somewhat higher, but it would not equal the population mean exactly. Similarly, if you took a second sample of 10 people from the same population, you would not expect the mean of this second sample to equal the mean of the first sample.

Recall that inferential statistics concern generalizing from a sample to a population. A critical part of inferential statistics involves determining how far sample statistics are likely to vary from each other and from the population **parameter**. (In this example, the sample statistics are the sample means and the population parameter is the population mean.) As the later portions of this chapter show, these determinations are based on sampling distributions.

6.1.1 Discrete Distributions

We will illustrate the concept of sampling distributions with a simple example. Figure 6.1 shows three pool balls, each with a number on it. Two of the balls are selected randomly (with replacement) and the average of their numbers is computed.

Figure 6.1: The pool balls.

All possible outcomes are shown below in Table 6.1.

Table 6.1: All possible outcomes when two balls are sampled with replacement.

Outcome	**Ball 1**	**Ball 2**	**Mean**
1	1	1	1.0
2	1	2	1.5
3	1	3	2.0
4	2	1	1.5
5	2	2	2.0
6	2	3	2.5
7	3	1	2.0
8	3	2	2.5
9	3	3	3.0

Notice that all the means are either 1.0, 1.5, 2.0, 2.5, or 3.0. The frequencies of these means are shown in Table 6-2. The relative frequencies are equal to the frequencies divided by nine because there are nine possible outcomes.

Table 6.2: Frequencies of means for N = 2.

Mean	**Frequency**	**Relative Frequency**
1.0	1	0.111
1.5	2	0.222
2.0	3	0.333
2.5	2	0.222
3.0	1	0.111

Figure 6.2 shows a relative frequency distribution of the means based on Table 6.2. This distribution is also a probability distribution since the Y-axis is the probability of obtaining a given mean from a sample of two balls in addition to being the relative frequency.

The distribution shown in Figure 6.2 is called the sampling distribution of the mean. Specifically, it is the sampling distribution of the mean for a sample size of 2 (N = 2). For this simple example, the distribution of pool balls and the sampling distribution are both discrete distributions. The pool balls have only the values 1, 2, and 3, and a sample mean can have one of only five values shown in Table 6.2.

There is an alternative way of conceptualizing a sampling distribution that will be useful for more complex distributions. Imagine that two balls are sampled (with replacement) and the mean of the two balls is computed and recorded. Then this process is repeated for a second sample, a third sample, and eventually thousands of samples. After thousands of samples are taken and the mean computed for each, a relative frequency distribution is drawn. The more samples, the closer the relative frequency distribution will come to the sampling distribution

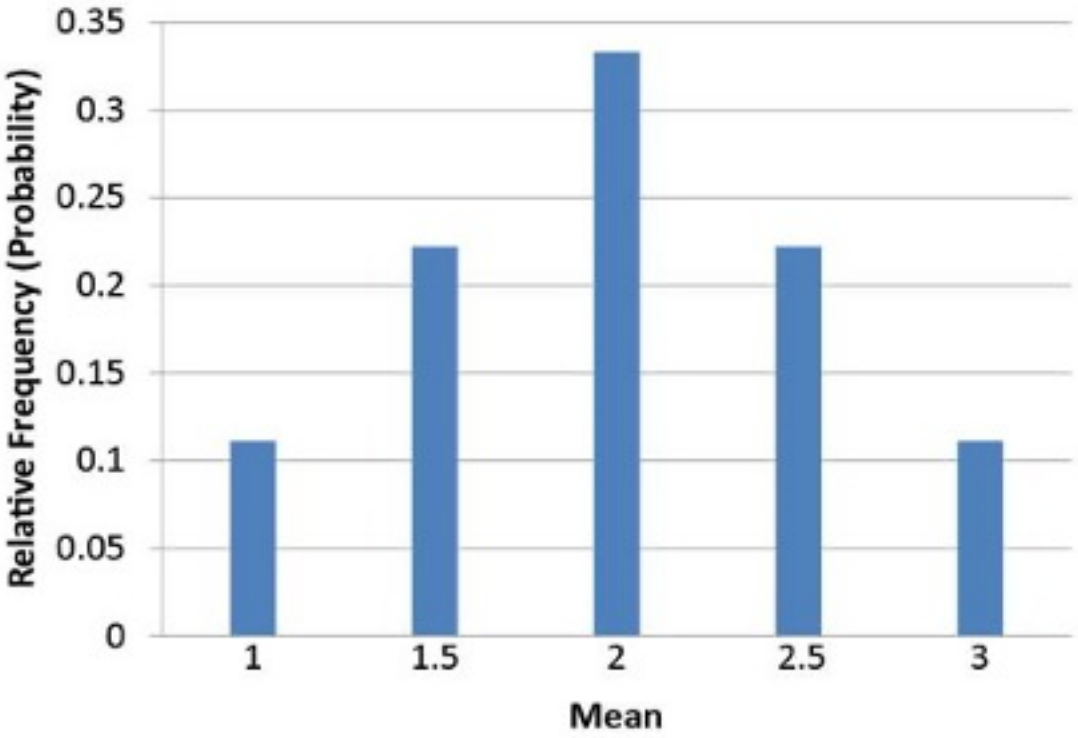


Figure 6.2: Distribution of means for N = 2.

shown in Figure 6.2. As the number of samples approaches infinity, the relative frequency distribution will approach the sampling distribution. This means that you can conceive of a sampling distribution as being a relative frequency distribution based on a very large number of samples. To be strictly correct, the relative frequency distribution approaches the sampling distribution as the number of samples approaches infinity.

It is important to keep in mind that every statistic, not just the mean, has a sampling distribution. For example, Table 6.3 shows all possible outcomes for the range of two numbers (larger number minus the smaller number). Table 6.4 shows the frequencies for each of the possible ranges and Figure 6.3 shows the sampling distribution of the range.

Table 6.3: All possible outcomes when two balls are sampled with replacement.

Outcome	**Ball 1**	**Ball 2**	**Range**
1	1	1	0
2	1	2	1
3	1	3	2
4	2	1	1
5	2	2	0
6	2	3	1
7	3	1	2
8	3	2	1
9	3	3	0

Table 6.4: Distribution of ranges for N = 2.

Range	**Frequency**	**Relative Frequency**
0	3	0.333
1	4	0.444
2	2	0.222

It is also important to keep in mind that there is a sampling distribution for various sample sizes. For simplicity, we have been using N = 2. The sampling distribution of the range for N = 3 is shown in Figure 6.4.

6.1.2 Continuous Distributions

In the previous section, the population consisted of three pool balls. Now we will consider sampling distributions when the population distribution is continuous. What if we had a thousand pool balls with numbers ranging from 0.001 to 1.000 in equal steps? (Although this distribution is not really continuous, it is close enough to be considered continuous for

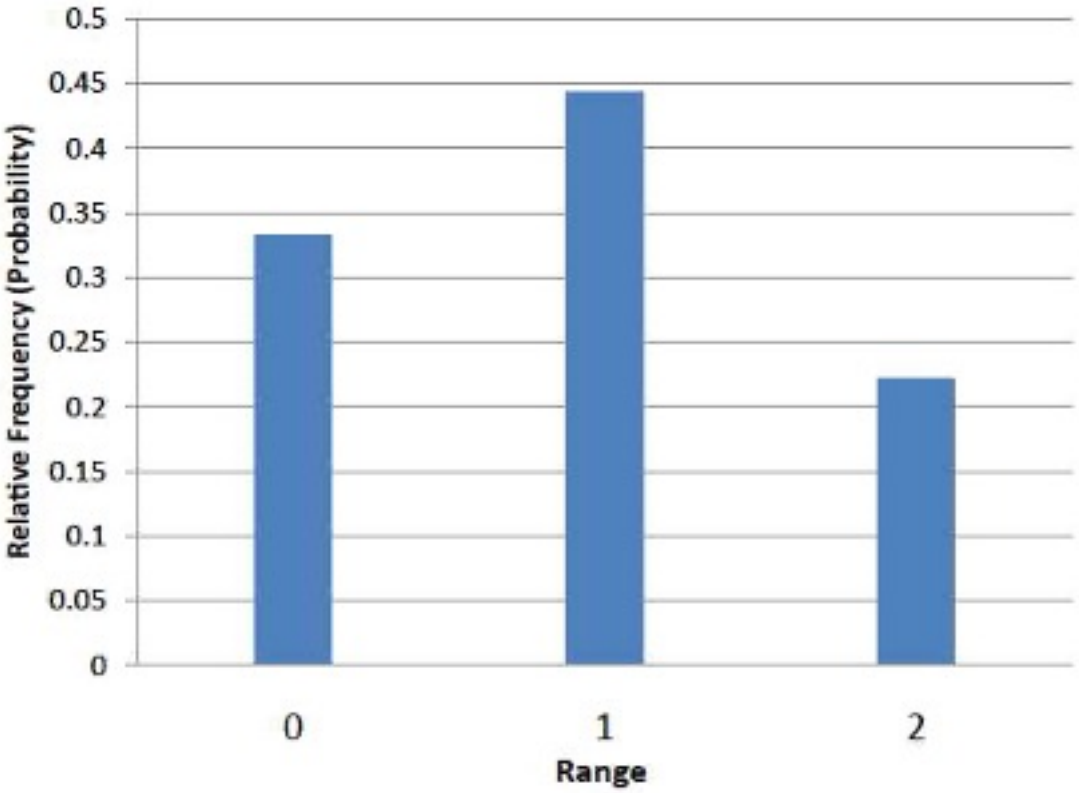


Figure 6.3: Distribution of ranges for N = 2.

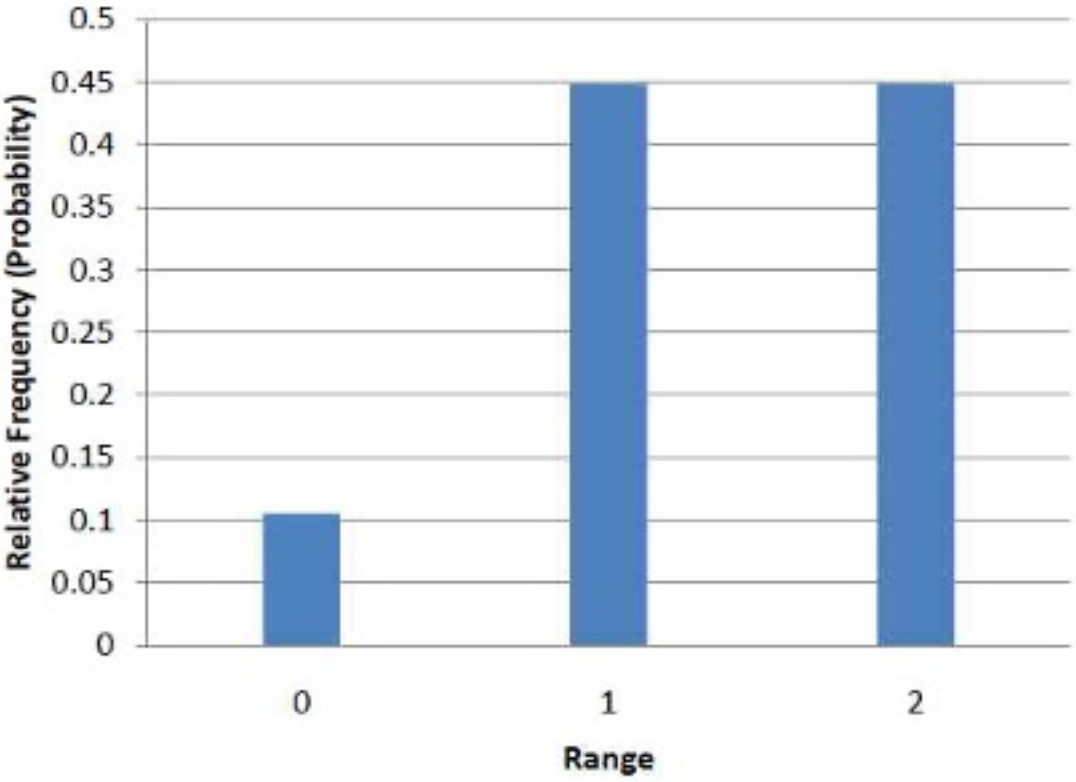


Figure 6.4: Distribution of ranges for N = 3.

practical purposes.) As before, we are interested in the distribution of means we would get if we sampled two balls and computed the mean of these two balls. In the previous example, we started by computing the mean for each of the nine possible outcomes. This would get a bit tedious for this example since there are 1,000,000 possible outcomes (1,000 for the first ball x 1,000 for the second). Therefore, it is more convenient to use our second conceptualization of sampling distributions which conceives of sampling distributions in terms of relative frequency distributions. Specifically, the relative frequency distribution that would occur if samples of two balls were repeatedly taken and the mean of each sample computed.

When we have a truly continuous distribution, it is not only impractical but actually impossible to enumerate all possible outcomes. Moreover, in continuous distributions, the probability of obtaining any single value is zero. Therefore, these values are called probability densities rather than probabilities.

6.1.3 Sampling Distributions and Inferential Statistics

As we stated in the beginning of this chapter, **sampling distributions** are important for inferential statistics. In the examples given so far, a population was specified and the sampling distribution of the mean and the range were determined. In practice, the process proceeds the other way: you collect sample data and from these data you estimate parameters of the sampling distribution. This knowledge of the sampling distribution can be very useful. For example, knowing the degree to which means from different samples would differ from each other and from the population mean would give you a sense of how close your particular sample mean is likely to be to the population mean. Fortunately, this information is directly available from a sampling distribution. The most common measure of how much sample means differ from each other is the standard deviation of the sampling distribution of the mean. This standard deviation is called the standard error of the mean. If all the sample means were very close to the population mean, then the standard error of the mean would be small. On the other hand, if the sample means varied considerably, then the standard error of the mean would be large.

To be specific, assume your sample mean were 125 and you estimated that the standard error of the mean were 5 (using a method shown in a later section). If you had a normal distribution, then it would be likely that your sample mean would be within 10 units of the population mean since most of a normal distribution is within two standard deviations of the mean.

Keep in mind that all statistics have sampling distributions, not just the mean. In later sections we will be discussing the sampling distribution of the variance, the sampling distribution of the difference between means, and the sampling distribution of Pearson's correlation, among others.

6.2 Sampling Distribution of the Mean[2]

The sampling distribution of the mean was defined in the section introducing sampling distributions. This section reviews some important properties of the sampling distribution of the mean introduced in the demonstrations in this chapter.

6.2.1 Mean

The mean of the sampling distribution of the mean is the mean of the population from which the scores were sampled. Therefore, if a population has a mean , then the mean of the sampling distribution of the mean is also . The symbol $_{\mathrm{M}}$ is used to refer to the mean of the sampling distribution of the mean. Therefore, the formula for the mean of the sampling distribution of the mean can be written as:

$$\mu\ M\ = \mu$$

6.2.2 Variance

The variance of the sampling distribution of the mean is computed as follows:

$$\sigma_M^2 = \frac{\sigma^2}{N}$$

That is, the variance of the sampling distribution of the mean is the population variance divided by N, the sample size (the number of scores used to compute a mean).[3] Thus, the larger the sample size, the smaller the variance of the sampling distribution of the mean.

The **standard error** of the mean is the standard deviation of the sampling distribution of the mean. It is therefore the square root of the variance of the sampling distribution of the mean and can be written as:

The standard error is represented by a because it is a standard deviation. The subscript (M) indicates that the standard error in question is the standard error of the mean.

6.2.3 Central Limit Theorem

The central limit theorem states that:

> Given a population with a finite mean μ and a finite non-zero variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean of μ and a variance of σ^2/N as N, the sample size, increases.

The expressions for the mean and variance of the sampling distribution of the mean are not new or remarkable. What is remarkable is that regardless of the shape of the parent population, the sampling distribution of the mean approaches a normal distribution as N increases. Figure 6.5 shows the results of the simulation for N = 2 and N = 10. The parent population was a uniform distribution. You can see that the distribution for N = 2 is far from a normal distribution. Nonetheless, it does show that the scores are denser in the middle than in the tails. For N = 10 the distribution is quite close to a normal distribution. Notice that the means of the two distributions are the same, but that the spread of the distribution for N = 10 is smaller.

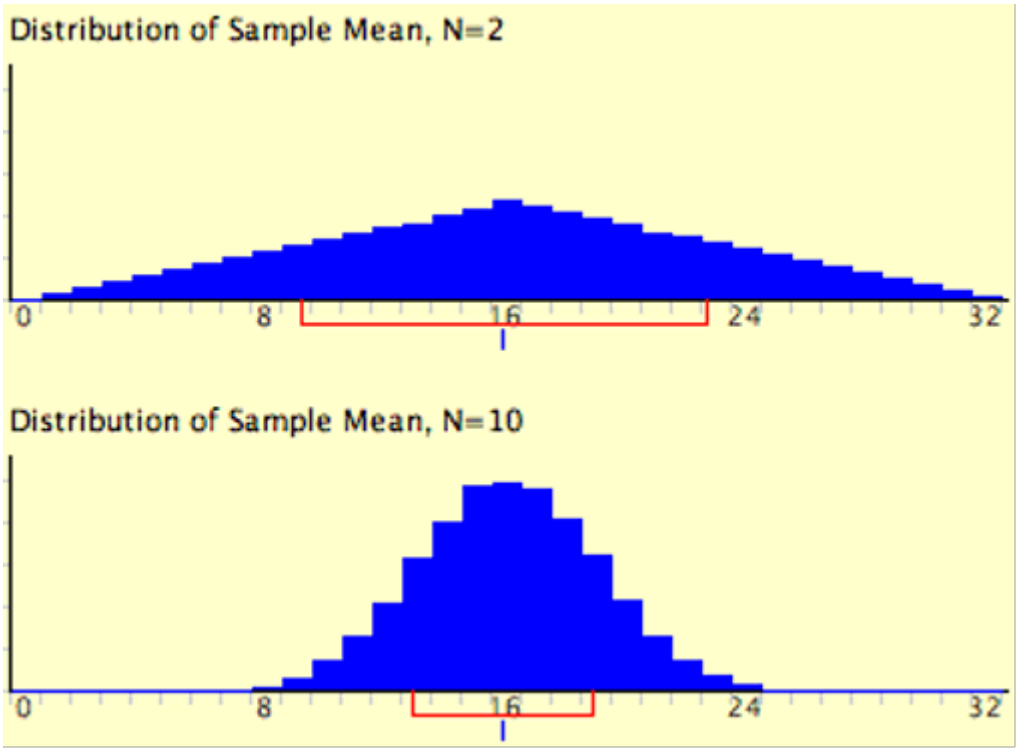


Figure 6.5: A simulation of a sampling distribution. The parent population is uniform. The blue line under "16" indicates that 16 is the mean. The red line extends from the mean plus and minus one standard deviation.

Figure 6.6 shows how closely the sampling distribution of the mean approximates a normal

distribution even when the parent population is very non-normal. If you look closely you can see that the sampling distributions do have a slight positive skew. The larger the sample size, the closer the sampling distribution of the mean would be to a normal distribution.

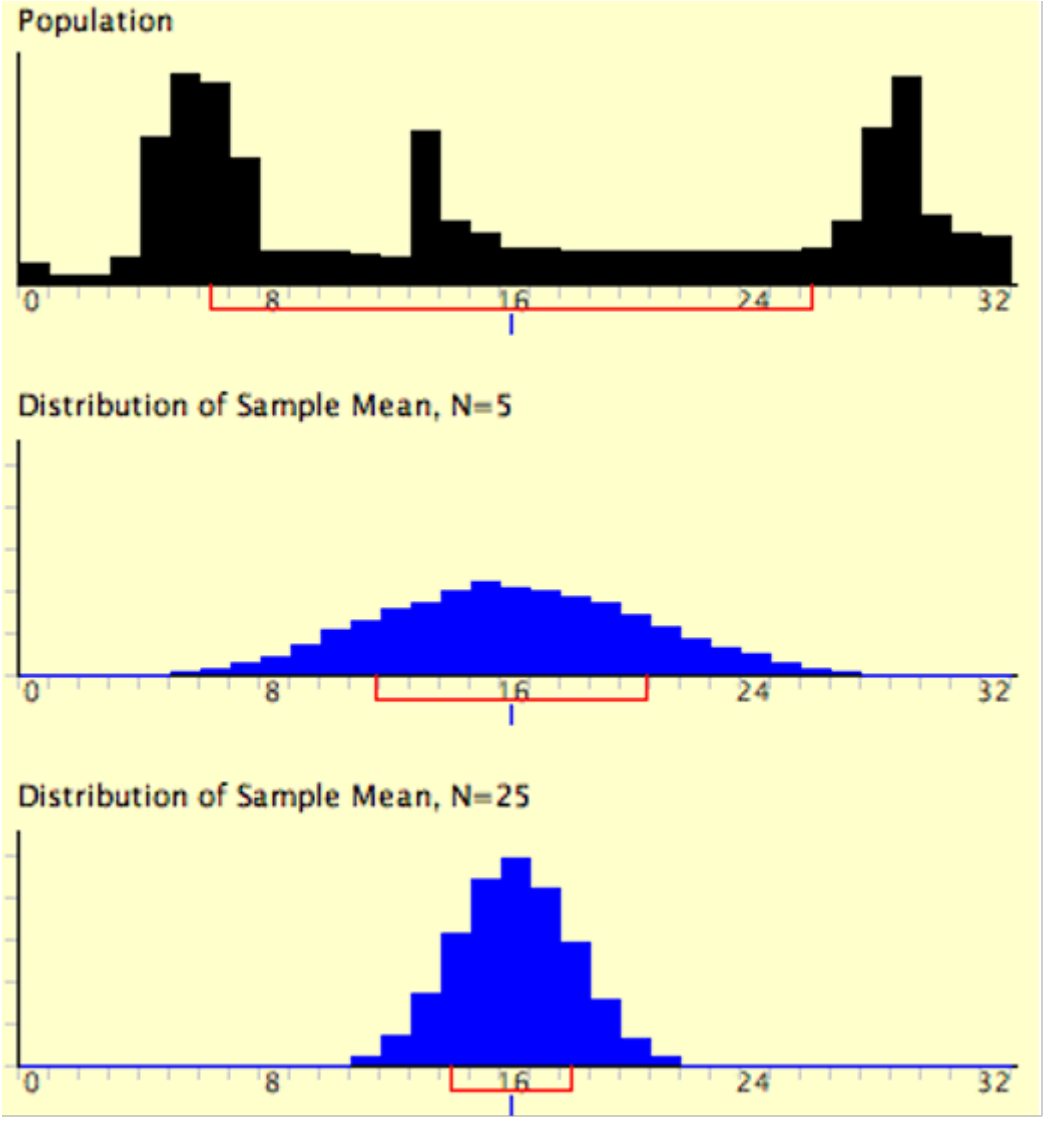


Figure 6.6: A simulation of a sampling distribution. The parent population is very non-normal.

6.3 Confidence Interval on the Mean[4]

When you compute a confidence interval on the mean, you compute the mean of a sample in order to estimate the mean of the population. Clearly, if you already knew the population mean,

there would be no need for a confidence interval. However, to explain how confidence intervals are constructed, we are going to work backwards and begin by assuming characteristics of the population. Then we will show how sample data can be used to construct a confidence interval.

Assume that the weights of 10-year-old children are normally distributed with a mean of 90 and a standard deviation of 36. What is the sampling distribution of the mean for a sample size of 9? Recall from the section on the sampling distribution of the mean that the mean of the sampling distribution is and the standard error of the mean is

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

For the present example, the sampling distribution of the mean has a mean of 90 and a standard deviation of $36/3 = 12$. Note that the standard deviation of a sampling distribution is its standard error. @ig-samplingdistn9 shows this distribution. The shaded area represents the middle 95% of the distribution and stretches from 66.48 to 113.52. These limits were computed by adding and subtracting 1.96 standard deviations to/from the mean of 90 as follows:

$$90 - (1.96)(12) = 66.48$$

$$90 + (1.96)(12) = 113.52$$

The value of 1.96 is based on the fact that 95% of the area of a normal distribution is within 1.96 standard deviations of the mean; 12 is the standard error of the mean.

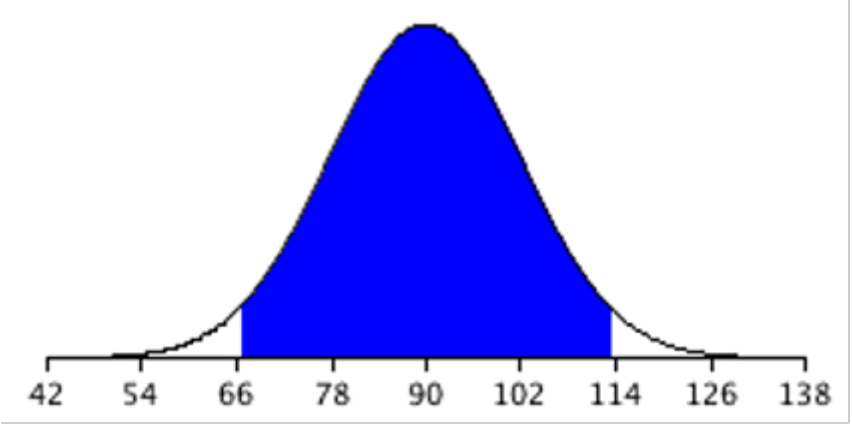


Figure 6.7: The sampling distribution of the mean for N=9. The middle 95% of the distribution is shaded.

Figure 6.7 shows that 95% of the means are no more than 23.52 units (1.96 standard deviations) from the mean of 90. Now consider the probability that a sample mean computed in a random sample is within 23.52 units of the population mean of 90. Since 95% of the distribution is within 23.52 of 90, the probability that the mean from any given sample will be within 23.52 of 90 is 0.95. This means that if we repeatedly compute the mean (M) from a sample, and create an interval ranging from M - 23.52 to M + 23.52, this interval will contain the population mean 95% of the time. In general, you compute the 95% confidence interval for the mean with the following formula:

$$\text{Lower limit} = M - Z\ .95\ \sigma\ M$$

$$\text{Upper limit} = M + Z\ .95\ \sigma\ M$$

where $Z_{.95}$ is the number of standard deviations extending from the mean of a normal distribution required to contain 0.95 of the area and $_{M}$ is the standard error of the mean.

If you look closely at this formula for a confidence interval, you will notice that you need to know the standard deviation ( ) in order to estimate the mean. This may sound unrealistic, and it is. However, computing a confidence interval when is known is easier than when has to be estimated, and serves a pedagogical purpose. Later in this section we will show how to compute a confidence interval for the mean when has to be estimated.

Suppose the following five numbers were sampled from a normal distribution with a standard deviation of 2.5: 2, 3, 5, 6, and 9. To compute the 95% confidence interval, start by computing the mean and standard error:

$$M = (2 + 3 + 5 + 6 + 9)/5 = 5.$$

$$\sigma_M = \frac{2.5}{\sqrt{5}} = 1.118.$$

$Z_{.95}$ can be found using the normal distribution calculator[5] and specifying that the shaded area is 0.95 and indicating that you want the area to be between the cutoff points. As shown in Figure 6.8, the value is 1.96. If you had wanted to compute the 99% confidence interval, you would have set the shaded area to 0.99 and the result would have been 2.58.

The confidence interval can then be computed as follows:

$$\text{Lower limit} = 5 - (1.96)(1.118) = 2.81$$

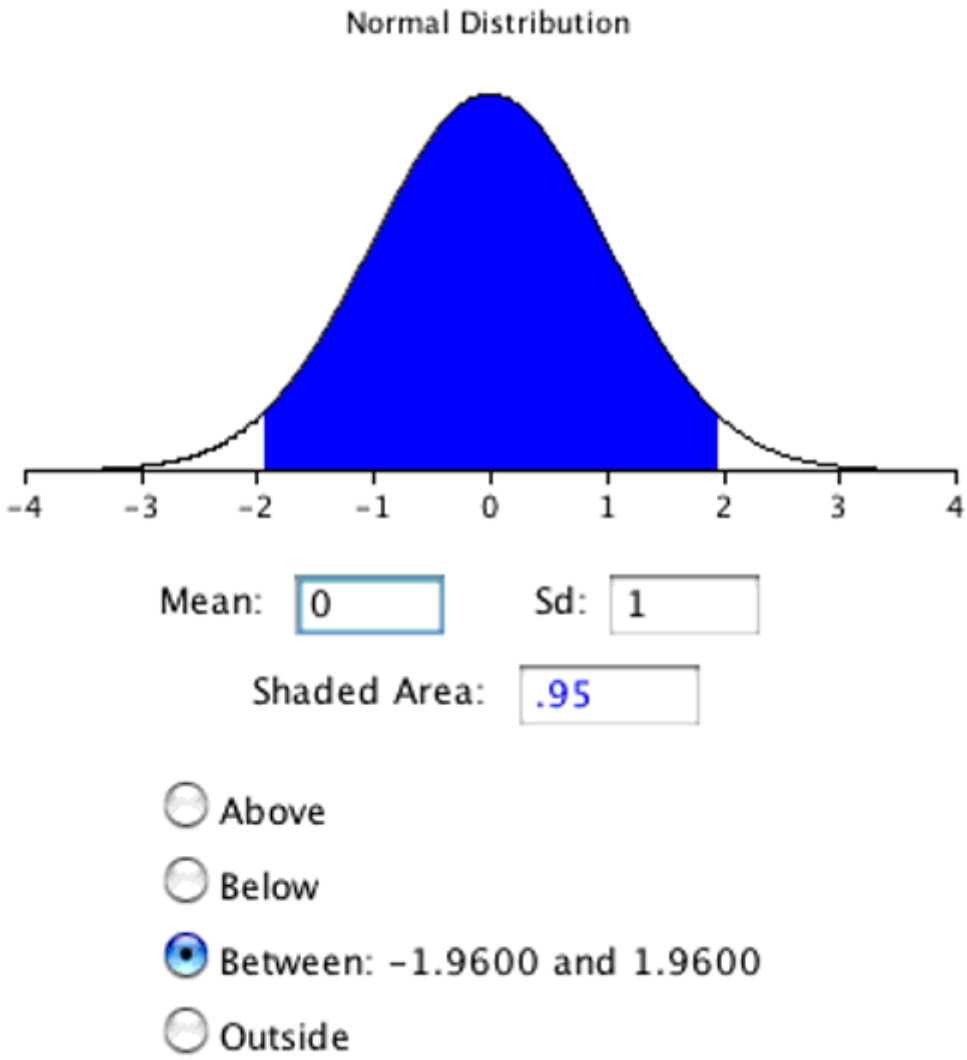


Figure 6.8: 95% of the area is between -1.96 and 1.96.

$$\text{Upper limit} = 5 + (1.96)(1.118) = 7.19$$

You should use the t distribution rather than the normal distribution when the variance is not known and has to be estimated from sample data. You will learn more about the t distribution in the next section. When the sample size is large, say 100 or above, the t distribution is very similar to the standard normal distribution. However, with smaller sample sizes, the t distribution has relatively more scores in its tails than does the normal distribution. As a result, you have to extend farther from the mean to contain a given proportion of the area. Recall that with a normal distribution, 95% of the distribution is within 1.96 standard deviations of the mean. Using the t distribution, if you have a sample size of only 5, 95% of the area is within 2.78 standard deviations of the mean. Therefore, the standard error of the mean would be multiplied by 2.78 rather than 1.96.

The values of t to be used in a confidence interval can be looked up in a table of the t distribution. A small version of such a table is shown in Table 6.5. The first column, df, stands for degrees of freedom, and for confidence intervals on the mean, df is equal to N - 1, where N is the sample size.

Table 6.5: Abbreviated t table.

df	**0.95**	**0.99**
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

You can also use an "inverse t distribution" calculator[6] to find the t values to use in confidence intervals.

Assume that the following five numbers are sampled from a normal distribution: 2, 3, 5, 6, and 9 and that the standard deviation is not known. The first steps are to compute the sample mean and variance:

$$M = 5$$

$$s2 = 7.5$$

The next step is to estimate the standard error of the mean. If we knew the population variance, we could use the following formula:

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

Instead we compute an estimate of the standard error (s_M):

$$s_M = \frac{s}{\sqrt{N}} = 1.225$$

The next step is to find the value of t. As you can see from Table 6.5, the value for the 95% interval for df = N - 1 = 4 is 2.776. The confidence interval is then computed just as it is when $_M$. The only differences are that s_M and t rather than $_M$ and Z are used.

$$\text{Lower limit} = 5 - (2.776)(1.225) = 1.60$$

$$\text{Upper limit} = 5 + (2.776)(1.225) = 8.40$$

More generally, the formula for the 95% confidence interval on the mean is:

$$\text{Lower limit} = M - (t_{CL})(s_M)$$

$$\text{Upper limit} = M + (t_{CL})(s_M)$$

where M is the sample mean, t_{CL} is the t for the confidence level desired (0.95 in the above example), and s_M is the estimated standard error of the mean.

We will finish with an analysis of the Stroop Data.[7] Specifically, we will compute a confidence interval on the mean difference score. Recall that 47 subjects named the color of ink that words were written in. The names conflicted so that, for example, they would name the ink color of the word "blue" written in red ink. The correct response is to say "red" and ignore the fact that the word is "blue." In a second condition, subjects named the ink color of colored rectangles.

Table 6.6: Response times in seconds for 10 subjects.

Naming Colored Rectangle	**Interference**	**Difference**
17	38	21
15	58	43
18	35	17
20	39	19
18	33	15
20	32	12
20	45	25
19	52	33
17	31	14
21	29	8

Table 6.6 shows the time difference between the interference and color-naming conditions for 10 of the 47 subjects. The mean time difference for all 47 subjects is 16.362 seconds and the standard deviation is 7.470 seconds. The standard error of the mean is 1.090. A t table shows the critical value of t for 47 - 1 = 46 degrees of freedom is 2.013 (for a 95% confidence interval). Therefore the confidence interval is computed as follows:

$$\text{Lower limit} = 16.362 - (2.013)(1.090) = 14.17$$

$$\text{Upper limit} = 16.362 + (2.013)(1.090) = 18.56$$

Therefore, the interference effect (difference) for the whole population is likely to be between 14.17 and 18.56 seconds.

6.4 The T Distribution[8]

In the introduction to normal distributions it was shown that 95% of the area of a normal distribution is within 1.96 standard deviations of the mean. Therefore, if you randomly sampled a value from a normal distribution with a mean of 100, the probability it would be within 1.96 of 100 is 0.95. Similarly, if you sample N values from the population, the probability that the sample mean (M) will be within 1.96 $_{M}$ of 100 is 0.95.

Now consider the case in which you have a normal distribution but you do not know the standard deviation. You sample N values and compute the sample mean (M) and estimate the

standard error of the mean ($_M$) with s_M. What is the probability that M will be within 1.96 s_M of the population mean ( )? This is a difficult problem because there are two ways in which M could be more than 1.96 s_M from : (1) M could, by chance, be either very high or very low and (2) s_M could, by chance, be very low. Intuitively, it makes sense that the probability of being within 1.96 standard errors of the mean should be smaller than in the case when the standard deviation is known (and cannot be underestimated). But exactly how much smaller? Fortunately, the way to work out this type of problem was solved in the early 20th century by W. S. Gosset who determined the distribution of a mean divided by an estimate of its standard error. This distribution is called the *Student's t distribution* or sometimes just the t distribution. Gosset worked out the t distribution and associated statistical tests while working for a brewery in Ireland. Because of a contractual agreement with the brewery, he published the article under the pseudonym "Student." That is why the t test is called the "Student's t test."

The **t distribution** is very similar to the normal distribution when the estimate of variance is based on a large sample, but the t distribution has relatively more scores in its tails when there is a small sample. When working with the t distribution, sample size is expressed in what are called degrees of freedom. Degrees of freedom will be discussed in more detail at the end of this chapter, but if we are estimating the standard error for a sample mean estimate, the degrees of freedom is simply equal to the sample size minus one (N-1).

Figure 6.9 shows t distributions with 2, 4, and 10 degrees of freedom and the standard normal distribution. Notice that the normal distribution has relatively more scores in the center of the distribution and the t distribution has relatively more in the tails. The t distribution approaches the normal distribution as the degrees of freedom increase.

Since the t distribution has more area in the tails, the percentage of the distribution within 1.96 standard deviations of the mean is less than the 95% for the normal distribution. Table 6.7 shows the number of standard deviations from the mean required to contain 95% and 99% of the area of the t distribution for various degrees of freedom. These are the values of t that you use in a confidence interval. The corresponding values for the normal distribution are 1.96 and 2.58 respectively. Notice that with few degrees of freedom, the values of t are much higher than the corresponding values for a normal distribution and that the difference decreases as the degrees of freedom increase. The values shown in Table 6-7 can be obtained from statistical software or an online calculator.[9]

Table 6.7: Abbreviated t table.

df	**0.95**	**0.99**
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355

10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

Returning to the problem posed at the beginning of this section, suppose you sampled 9 values from a normal population and estimated the standard error of the mean ($_M$) with s_M. What is the probability that M would be within $1.96s_M$ of ? Since the sample size is 9, there are N - 1 = 8 df. From Table 6.7, you can see that with 8 df the probability is 0.95 that the mean will be within 2.306 s_M of . The probability that it will be within 1.96 s_M of is therefore lower than 0.95.

As shown in Figure 6.10, a t distribution calculator[10] can be used to find that 0.086 of the area of a t distribution is more than 1.96 standard deviations from the mean, so the probability that M would be less than 1.96sM from is 1 - 0.086 = 0.914.

As expected, this probability is less than 0.95 that would have been obtained if M had been known instead of estimated.

6.5 Degrees of Freedom[11]

Some estimates are based on more information than others. For example, an estimate of the variance based on a sample size of 100 is based on more information than an estimate of the variance based on a sample size of 5. The **degrees of freedom (df)** of an estimate is the number of independent pieces of information on which the estimate is based.

As an example, let's say that we know that the mean height of Martians is 6 and wish to estimate the variance of their heights. We randomly sample one Martian and find that its height is 8. Recall that the variance is defined as the mean squared deviation of the values from their population mean. We can compute the squared deviation of our value of 8 from the population mean of 6 to find a single squared deviation from the mean. This single squared deviation from the mean, $(8\text{-}6)^2 = 4$, is an estimate of the mean squared deviation for all Martians. Therefore, based on this sample of one, we would estimate that the population variance is 4. This estimate is based on a single piece of information and therefore has 1 df. If we sampled another Martian and obtained a height of 5, then we could compute a second estimate of the variance, $(5\text{-}6)^2 = 1$. We could then average our two estimates (4 and 1) to obtain an estimate of 2.5. Since this estimate is based on two independent pieces of information, it has two degrees of freedom. The two estimates are independent because they are based on two independently and randomly selected Martians. The estimates would not be independent if after sampling one Martian, we decided to choose its brother as our second Martian.

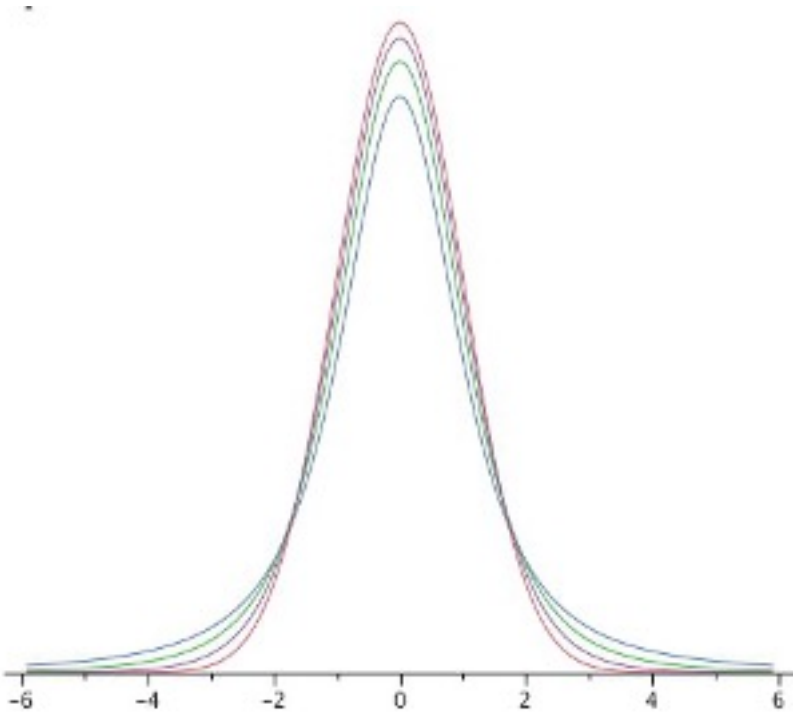


Figure 6.9: A comparison of t distributions with 2, 4, and 10 df and the standard normal distribution. The distribution with the lowest peak is the 2 df distribution, the next lowest is 4 df, the lowest after that is 10 df, and the highest is the standard normal distribution.

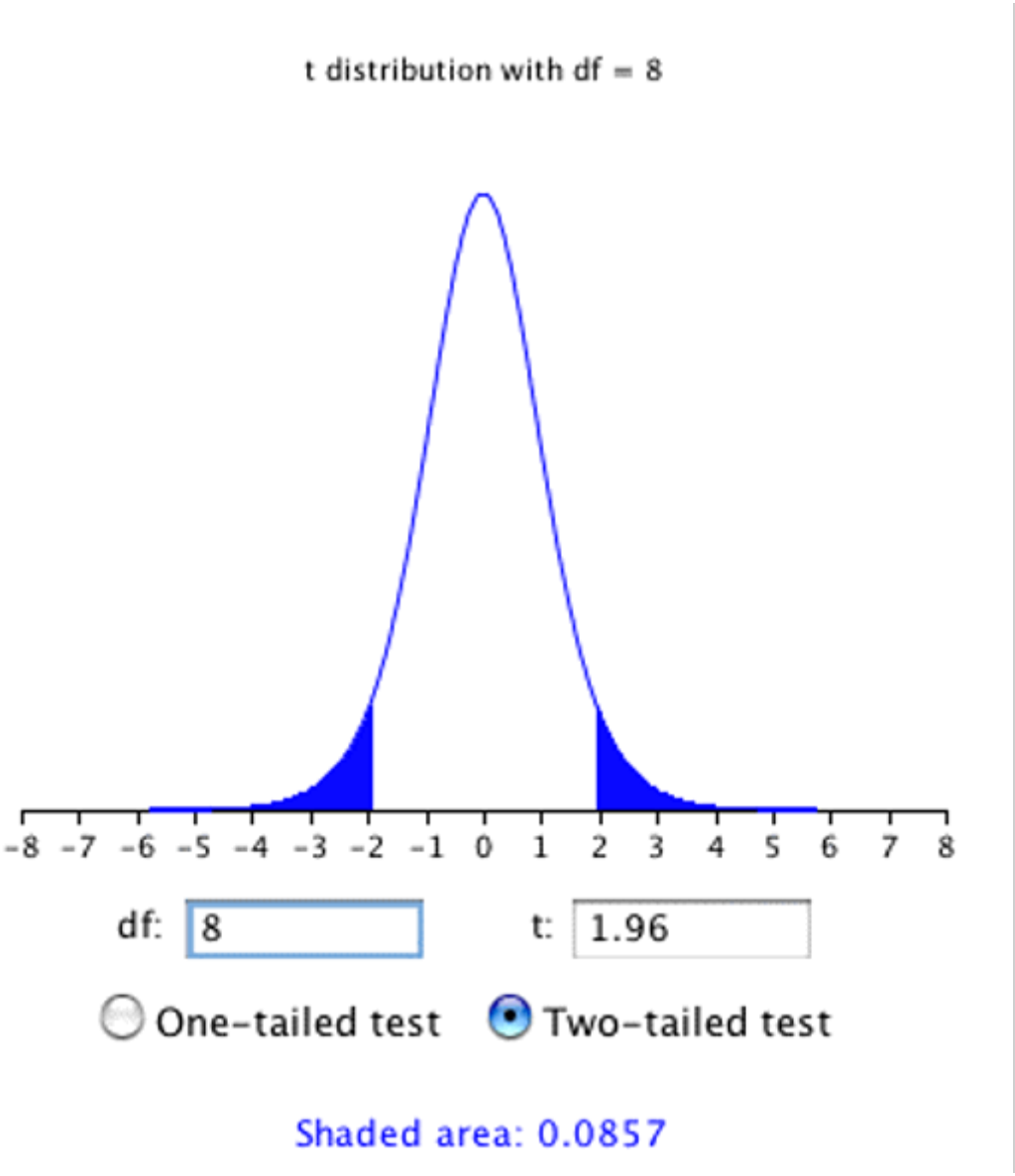


Figure 6.10: Area more than 1.96 standard deviations from the mean in a t distribution with 8 df. Note that the two-tailed button is selected so that the area in both tails will be included.

As you are probably thinking, it is pretty rare that we know the population mean when we are estimating the variance. Instead, we have to first estimate the population mean ( ) with the sample mean (M). The process of estimating the mean affects our degrees of freedom as shown below.

Returning to our problem of estimating the variance in Martian heights, let's assume we do not know the population mean and therefore we have to estimate it from the sample. We have sampled two Martians and found that their heights are 8 and 5. Therefore M, our estimate of the population mean, is

$$M = (8 + 5)/2 = 6.5.$$

We can now compute two estimates of variance:

$$\text{Estimate 1} = (8 - 6.5)^2 = 2.25$$

$$\text{Estimate 2} = (5 - 6.5)^2 = 2.25$$

Now for the key question: Are these two estimates independent? The answer is no because each height contributed to the calculation of M. Since the first Martian's height of 8 influenced M, it also influenced Estimate 2. If the first height had been, for example, 10, then M would have been 7.5 and Estimate 2 would have been $(5\text{-}7.5)^2 = 6.25$ instead of 2.25. The important point is that the two estimates are not independent and therefore we do not have two degrees of freedom. Another way to think about the non-independence is to consider that if you knew the mean and one of the scores, you would know the other score. For example, if one score is 5 and the mean is 6.5, you can compute that the total of the two scores is 13 and therefore that the other score must be 13-5 = 8.

In general, the degrees of freedom for an estimate is equal to the number of values minus the number of parameters estimated en route to the estimate in question. In the Martians example, there are two values (8 and 5) and we had to estimate one parameter ( ) on the way to estimating the parameter of interest (^{2}). Therefore, the estimate of variance has 2 - 1 = 1 degree of freedom. If we had sampled 12 Martians, then our estimate of variance would have had 11 degrees of freedom. Therefore, the degrees of freedom of an estimate of variance is equal to N - 1, where N is the number of observations.

Recall from the section on variability that the formula for estimating the variance in a sample is:

$$s^2 = \frac{\Sigma(X - M)^2}{N - 1}$$

The denominator of this formula is the degrees of freedom.

So far, we've only seen examples where the degrees of freedom is equal to N - 1. But in later chapters, we'll see examples of statistical inference tools that require estimating more than one parameter en route to the estimate in question, and therefore we'll need to subtract more than one from the number of observations to get the degrees of freedom. For example, the degrees of freedom might be calculated as N - 5 or N - 3, depending on what we are estimating.

---

[1] This subsection is adapted from David M. Lane. "Introduction to Sampling Distributions." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/sampling_distributions/intro_samp_dist.html

[2] This subsection is adapted from David M. Lane. "Sampling Distribution of the Mean." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/sampling_distributions/samp_dist_mean.html

[3] This expression can be derived very easily from the variance sum law. Let's begin by computing the variance of the sampling distribution of the sum of three numbers sampled from a population with variance ^{2}. The variance of the sum would be $^2 + ^2 + ^2$. For N numbers, the variance would be N ^{2}. Since the mean is 1/N times the sum, the variance of the sampling distribution of the mean would be $1/N^2$ times the variance of the sum, which equals $^2/N$.

[4] This subsection is adapted from David M. Lane. "Confidence Interval on the Mean." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/estimation/mean.html

[5] http://onlinestatbook.com/2/calculators/normal_dist.html

[6] http://onlinestatbook.com/2/calculators/inverse_t_dist.html

[7] http://onlinestatbook.com/2/case_studies/stroop.html

[8] This section is adapted from David M. Lane. "t Distribution." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/estimation/t_distribution.html

[9] http://onlinestatbook.com/2/calculators/inverse_t_dist.html

[10] http://onlinestatbook.com/2/calculators/t_dist.html

[11] This section is adapted from David M. Lane. "Degrees of Freedom." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/estimation/df.html

7 Hypothesis Testing

7.1 Introduction to Hypothesis Testing[1]

The statistician R. Fisher explained the concept of hypothesis testing with a story of a lady tasting tea. Here we will present an example based on James Bond who insisted that martinis should be shaken rather than stirred. Let's consider a hypothetical experiment to determine whether Mr. Bond can tell the difference between a shaken and a stirred martini. Suppose we gave Mr. Bond a series of 16 taste tests. In each test, we flipped a fair coin to determine whether to stir or shake the martini. Then we presented the martini to Mr. Bond and asked him to decide whether it was shaken or stirred. Let's say Mr. Bond was correct on 13 of the 16 taste tests. Does this prove that Mr. Bond has at least some ability to tell whether the martini was shaken or stirred?

This result does not prove that he does; it could be he was just lucky and guessed right 13 out of 16 times. But how plausible is the explanation that he was just lucky? To assess its plausibility, we determine the probability that someone who was just guessing would be correct 13/16 times or more. This probability can be computed from the binomial distribution, and a binomial distribution calculator[2] shows it to be 0.0106. This is a pretty low probability, and therefore someone would have to be very lucky to be correct 13 or more times out of 16 if they were just guessing. So either Mr. Bond was very lucky, or he can tell whether the drink was shaken or stirred. The hypothesis that he was guessing is not proven false, but considerable doubt is cast on it. Therefore, there is strong evidence that Mr. Bond can tell whether a drink was shaken or stirred.

Let's consider another example. The case study Physicians' Reactions[3] sought to determine whether physicians spend less time with obese patients. Physicians were sampled randomly and each was shown a chart of a patient complaining of a migraine headache. They were then asked to estimate how long they would spend with the patient. The charts were identical except that for half the charts, the patient was obese and for the other half, the patient was of average weight. The chart a particular physician viewed was determined randomly. Thirty-three physicians viewed charts of average-weight patients and 38 physicians viewed charts of obese patients.

The mean time physicians reported that they would spend with obese patients was 24.7 minutes as compared to a mean of 31.4 minutes for average-weight patients. How might this difference between means have occurred? One possibility is that physicians were influenced by the weight of the patients. On the other hand, perhaps by chance, the physicians who viewed charts of the

obese patients tend to see patients for less time than the other physicians. Random assignment of charts does not ensure that the groups will be equal in all respects other than the chart they viewed. In fact, it is certain the two groups differed in many ways by chance. The two groups could not have exactly the same mean age (if measured precisely enough such as in days). Perhaps a physician's age affects how long physicians see patients. There are innumerable differences between the groups that could affect how long they view patients. With this in mind, is it plausible that these chance differences are responsible for the difference in times?

To assess the plausibility of the hypothesis that the difference in mean times is due to chance, we compute the probability of getting a difference as large or larger than the observed difference (31.4 - 24.7 = 6.7 minutes) if the difference were, in fact, due solely to chance. Using methods presented in a later chapter, this probability can be computed to be 0.0057. Since this is such a low probability, we have confidence that the difference in times is due to the patient's weight and is not due to chance.

7.1.1 The Probability Value

It is very important to understand precisely what the probability values mean. In the James Bond example, the computed probability of 0.0106 is the probability he would be correct on 13 or more taste tests (out of 16) if he were just guessing.

It is easy to mistake this probability of 0.0106 as the probability he cannot tell the difference. This is not at all what it means.

The probability of 0.0106 is the probability of a certain outcome (13 or more out of 16) assuming a certain state of the world (James Bond was only guessing). It is not the probability that a state of the world is true. Although this might seem like a distinction without a difference, consider the following example. An animal trainer claims that a trained bird can determine whether or not numbers are evenly divisible by 7. In an experiment assessing this claim, the bird is given a series of 16 test trials. On each trial, a number is displayed on a screen and the bird pecks at one of two keys to indicate its choice. The numbers are chosen in such a way that the probability of any number being evenly divisible by 7 is 0.50. The bird is correct on 9/16 choices. Using the binomial calculator, we can compute that the probability of being correct nine or more times out of 16 if one is only guessing is 0.40. Since a bird who is only guessing would do this well 40% of the time, these data do not provide convincing evidence that the bird can tell the difference between the two types of numbers. As a scientist, you would be very skeptical that the bird had this ability. Would you conclude that there is a 0.40 probability that the bird can tell the difference? Certainly not! You would think the probability is much lower than 0.0001.

To reiterate, the **probability value (p value)** is the probability of an outcome (9/16 or better) and not the probability of a particular state of the world (the bird was only guessing). In statistics, it is conventional to refer to possible states of the world as hypotheses since they are hypothesized states of the world. Using this terminology, the probability value is the

probability of an outcome given the hypothesis. It is not the probability of the hypothesis given the outcome.

This is not to say that we ignore the probability of the hypothesis. If the probability of the outcome given the hypothesis is sufficiently low, we have evidence that the hypothesis is false. However, we do not compute the probability that the hypothesis is false. In the James Bond example, the hypothesis is that he cannot tell the difference between shaken and stirred martinis. The probability value is low (0.0106), thus providing evidence that he can tell the difference. However, we have not computed the probability that he can tell the difference. A branch of statistics called Bayesian statistics provides methods for computing the probabilities of hypotheses. These computations require that one specify the probability of the hypothesis before the data are considered and, therefore, are difficult to apply in some contexts.

7.1.2 The Null Hypothesis

The hypothesis that an apparent effect is due to chance is called the **null hypothesis**. In the Physicians' Reactions example, the null hypothesis is that in the population of physicians, the mean time expected to be spent with obese patients is equal to the mean time expected to be spent with average-weight patients. This null hypothesis can be written as:

$$\mu_{obese} = \mu_{average}$$

or as

$$\mu_{obese} - \mu_{average} = 0.$$

The null hypothesis in a correlational study of the relationship between high school grades and college grades would typically be that the population correlation is 0. This can be written as

$$\rho = 0$$

where ρ is the population correlation (not to be confused with r, the correlation in the sample).

Although the null hypothesis is usually that the value of a *population parameter* is 0, there are occasions in which the null hypothesis is a value other than 0. For example, if one were testing whether a subject differed from chance in their ability to determine whether a flipped coin would come up heads or tails, the null hypothesis would be that $\pi = 0.5$.

Keep in mind that the null hypothesis is typically the opposite of the researcher's hypothesis. In the Physicians' Reactions study, the researchers hypothesized that physicians would expect

to spend less time with obese patients. The null hypothesis that the two types of patients are treated identically is put forward with the hope that it can be discredited and therefore rejected. If the null hypothesis were true, a difference as large or larger than the sample difference of 6.7 minutes would be very unlikely to occur. Therefore, the researchers rejected the null hypothesis of no difference and concluded that in the population, physicians intend to spend less time with obese patients.

If the null hypothesis is rejected, then the alternative to the null hypothesis (called the alternative hypothesis) is accepted. The **alternative hypothesis** is simply the reverse of the null hypothesis. If the null hypothesis

$$\mu_{obese} = \mu_{average}$$

is rejected, then there are two alternatives:

$$\mu_{obese} \leq \mu_{average}$$

$$\mu_{obese} \geq \mu_{average}$$

Naturally, the direction of the sample means determines which alternative is adopted. Some textbooks have incorrectly argued that rejecting the null hypothesis that two population means are equal does not justify a conclusion about which population mean is larger. Kaiser (1960)[4] showed how it is justified to draw a conclusion about the direction of the difference.

7.2 Steps in Hypothesis Testing[5]

There's much to learn about hypothesis testing, but before going any further, here's an overview of the four basic steps of any hypothesis test. Some of the details won't make sense yet, but we'll explain them in more detail in the following sections.

1. The first step is to ***specify the null hypothesis***. For a two-tailed test, the null hypothesis is typically that a parameter equals zero although there are exceptions. A typical null hypothesis is μ_1 - $\mu_2 = 0$ which is equivalent to $\mu_1 = \mu_2$. For a one-tailed test, the null hypothesis is either that a parameter is greater than or equal to zero or that a parameter is less than or equal to zero. If the prediction is that μ_1 is larger than μ_2, then the null hypothesis (the reverse of the prediction) is μ_2 - μ_1 0. This is equivalent to μ_1 μ_2.
2. The second step is to ***specify the*** α level which is also known as the significance level. Typical values are 0.05 and 0.01.

3. The third step is to ***compute the probability value*** (also known as the p value). This is the probability of obtaining a sample statistic as different or more different from the parameter specified in the null hypothesis given that the null hypothesis is true.

4. Finally, ***compare the probability value with the*** α level. If the probability value is lower then you reject the null hypothesis. Keep in mind that rejecting the null hypothesis is not an all-or-none decision. The lower the probability value, the more confidence you can have that the null hypothesis is false. However, if your probability value is higher than the conventional α level of 0.05, most scientists will consider your findings inconclusive. Failure to reject the null hypothesis does not constitute support for the null hypothesis. It just means you do not have sufficiently strong data to reject it.

7.3 One- and Two-Tailed Tests[6]

In the James Bond case study,[7] Mr. Bond was given 16 trials on which he judged whether a martini had been shaken or stirred. He was correct on 13 of the trials. From the binomial distribution, we know that the probability of being correct 13 or more times out of 16 if one is only guessing is 0.0106. Figure 7.1 shows a graph of the binomial distribution. The red bars show the values greater than or equal to 13. As you can see in the figure, the probabilities are calculated for the upper tail of the distribution. A probability calculated in only one tail of the distribution is called a "one-tailed probability."

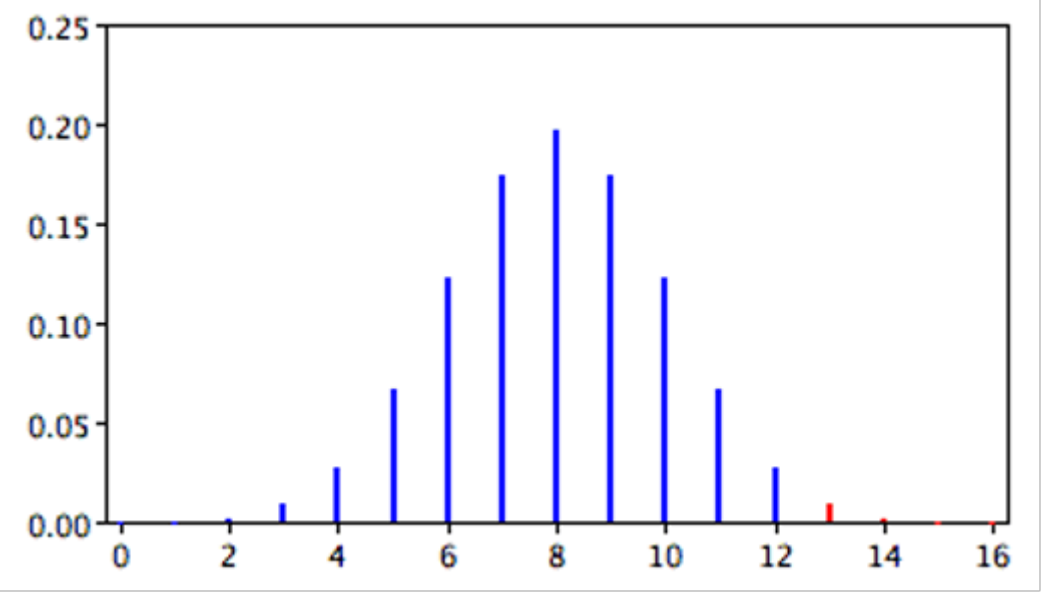


Figure 7.1: The binomial distribution. The upper (right-hand) tail is red.

A slightly different question can be asked of the data: "What is the probability of getting a result as extreme or more extreme than the one observed?" Since the chance expectation is 8/16, a result of 3/16 is equally as extreme as 13/16. Thus, to calculate this probability, we would consider both tails of the distribution. Since the binomial distribution is symmetric when $\pi = 0.5$, this probability is exactly double the probability of 0.0106 computed previously. Therefore, p = 0.0212. A probability calculated in both tails of a distribution is called a "two-tailed probability" (see Figure 7.2).

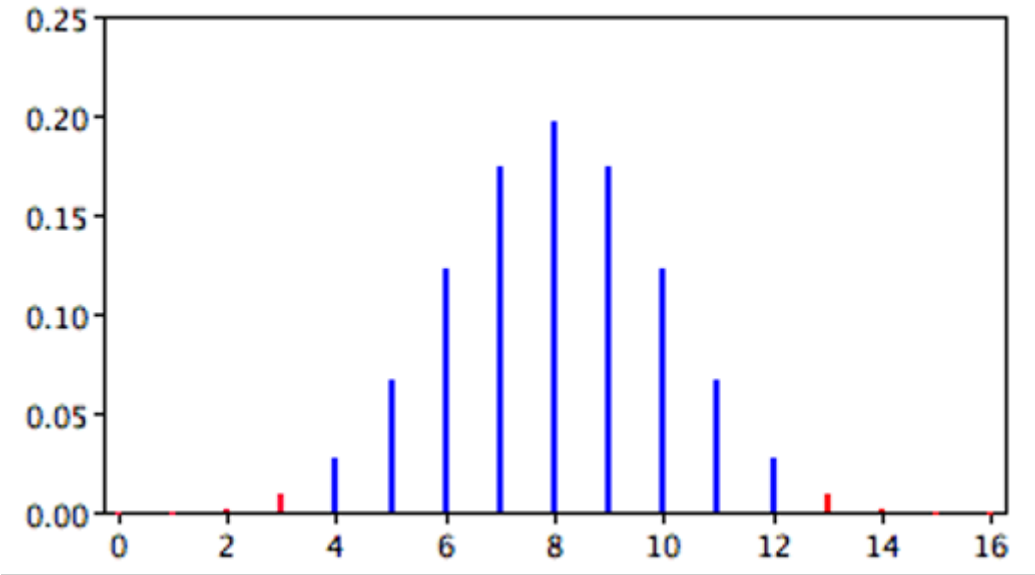


Figure 7.2: The binomial distribution. Both tails are red.

Should the one-tailed or the two-tailed probability be used to assess Mr. Bond's performance? That depends on the way the question is posed. If we are asking whether Mr. Bond can tell the difference between shaken or stirred martinis, then we would conclude he could if he performed either much better than chance or much worse than chance. If he performed much worse than chance, we would conclude that he can tell the difference, but he does not know which is which. Therefore, since we are going to reject the null hypothesis if Mr. Bond does either very well or very poorly, we will use a two-tailed probability.

On the other hand, if our question is whether Mr. Bond is better than chance at determining whether a martini is shaken or stirred, we would use a one-tailed probability. What would the one-tailed probability be if Mr. Bond were correct on only 3 of the 16 trials? Since the one-tailed probability is the probability of the right-hand tail, it would be the probability of getting 3 or more correct out of 16. This is a very high probability and the null hypothesis would not be rejected.

The null hypothesis for the two-tailed test is $\pi = 0.5$. By contrast, the null hypothesis for the one-tailed test is $\pi \quad 0.5$. Accordingly, we reject the two-tailed hypothesis if the sample

proportion deviates greatly from 0.5 in either direction. The one-tailed hypothesis is rejected only if the sample proportion is much greater than 0.5. The alternative hypothesis in the two-tailed test is $\pi \neq 0.5$. In the one-tailed test it is $\pi > 0.5$.

You should always decide whether you are going to use a one-tailed or a two-tailed probability before looking at the data. Statistical tests that compute one-tailed probabilities are called one-tailed tests; those that compute two-tailed probabilities are called two-tailed tests. Two-tailed tests are much more common than one-tailed tests in scientific research because an outcome signifying that something other than chance is operating is usually worth noting. One-tailed tests are appropriate when it is not important to distinguish between no effect and an effect in the unexpected direction. For example, consider an experiment designed to test the efficacy of a treatment for the common cold. The researcher would only be interested in whether the treatment was better than a placebo control. It would not be worth distinguishing between the case in which the treatment was worse than a placebo and the case in which it was the same because in both cases the drug would be worthless.

Some have argued that a one-tailed test is justified whenever the researcher predicts the direction of an effect. The problem with this argument is that if the effect comes out strongly in the non-predicted direction, the researcher is not justified in concluding that the effect is not zero. Since this is unrealistic, one-tailed tests are usually viewed skeptically if justified on this basis alone.

7.4 Significance Testing[8]

A low probability value casts doubt on the null hypothesis. How low must the probability value be in order to conclude that the null hypothesis is false? Although there is clearly no right or wrong answer to this question, it is conventional to conclude the null hypothesis is false if the probability value is less than 0.05. More conservative researchers conclude the null hypothesis is false only if the probability value is less than 0.01. When a researcher concludes that the null hypothesis is false, the researcher is said to have rejected the null hypothesis. The probability value below which the null hypothesis is rejected is called the α (alpha) level or simply α. It is also called the significance level.

When the null hypothesis is rejected, the effect is said to be **statistically significant**. For example, in the Physicians' Reactions case study,[9] the probability value is 0.0057. Therefore, the effect of obesity is statistically significant and the null hypothesis that obesity makes no difference is rejected. It is very important to keep in mind that statistical significance means only that the null hypothesis of exactly no effect is rejected; it does not mean that the effect is important, which is what "significant" usually means. When an effect is significant, you can have confidence the effect is not exactly zero. Finding that an effect is significant does not tell you about how large or important the effect is.

Do not confuse statistical significance with practical significance. A small effect can be highly significant if the sample size is large enough.

Why does the word "significant" in the phrase "statistically significant" mean something so different from other uses of the word? Interestingly, this is because the meaning of "significant" in everyday language has changed. It turns out that when the procedures for hypothesis testing were developed, something was "significant" if it signified something. Thus, finding that an effect is statistically significant signifies that the effect is real and not due to chance. Over the years, the meaning of "significant" changed, leading to the potential misinterpretation.

There are two approaches (at least) to conducting significance tests. In one (favored by R. Fisher), a significance test is conducted and the probability value reflects the strength of the evidence against the null hypothesis. If the probability is below 0.01, the data provide strong evidence that the null hypothesis is false. If the probability value is below 0.05 but larger than 0.01, then the null hypothesis is typically rejected, but not with as much confidence as it would be if the probability value were below 0.01. Probability values between 0.05 and 0.10 provide weak evidence against the null hypothesis and, by convention, are not considered low enough to justify rejecting it. Higher probabilities provide less evidence that the null hypothesis is false.

The alternative approach (favored by the statisticians Neyman and Pearson) is to specify an α level before analyzing the data. If the data analysis results in a probability value below the α level, then the null hypothesis is rejected; if it is not, then the null hypothesis is not rejected. According to this perspective, if a result is significant, then it does not matter how significant it is. Moreover, if it is not significant, then it does not matter how close to being significant it is. Therefore, if the 0.05 level is being used, then probability values of 0.049 and 0.001 are treated identically. Similarly, probability values of 0.06 and 0.34 are treated identically.

The former approach (preferred by Fisher) is more suitable for scientific research and will be adopted here. The latter is more suitable for applications in which a yes/no decision must be made. For example, if a statistical analysis were undertaken to determine whether a machine in a manufacturing plant were malfunctioning, the statistical analysis would be used to determine whether or not the machine should be shut down for repair. The plant manager would be less interested in assessing the weight of the evidence than knowing what action should be taken. There is no need for an immediate decision in scientific research where a researcher may conclude that there is some evidence against the null hypothesis, but that more research is needed before a definitive conclusion can be drawn.

7.5 Testing a Single Mean[10]

The way we calculate the probability (p) value for a hypothesis test depends on what type of statement is made in our null hypothesis. Normally, statistical software will automatically compute a p value behind the scenes, but we still want to learn a bit about how the software

comes up with this value. To illustrate what these calculations can look like, this section will focus on what to do if we want to test a null hypothesis stating that the population mean is equal to some hypothesized value. For example, suppose an experimenter wanted to know if people are influenced by a subliminal message and performed the following experiment. Each of nine subjects is presented with a series of 100 pairs of pictures. As a pair of pictures is presented, a subliminal message is presented suggesting the picture that the subject should choose. The question is whether the (population) mean number of times the suggested picture is chosen is equal to 50. In other words, the null hypothesis is that the population mean (μ) is 50. The (hypothetical) data are shown in Table 7.1. The data in Table 7.1 have a sample mean (M) of 51. Thus the sample mean differs from the hypothesized population mean by 1.

Table 7.1: Distribution of scores.

Frequency
45
48
49
49
51
52
53
55
57

The significance test consists of computing the probability of a sample mean differing from μ by one (the difference between the hypothesized population mean and the sample mean) or more. The first step is to determine the sampling distribution of the mean. As we learned in the prior chapter, the mean and standard deviation of the sampling distribution of the mean are

$$\mu_M = \mu$$

and

$$\sigma_M = \frac{\sigma}{\sqrt{(N)}}$$

respectively. It is clear that $\mu_M = 50$. In order to compute the standard deviation of the sampling distribution of the mean, we have to know the population standard deviation (σ).

The current example was constructed to be one of the few instances in which the standard deviation is known. In practice, it is very unlikely that you would know σ and therefore you

would use s, the sample estimate of σ. However, it is instructive to see how the probability is computed if σ is known before proceeding to see how it is calculated when σ is estimated.

For the current example, if the null hypothesis is true, then based on the binomial distribution, one can compute that variance of the number correct is

$$\sigma^2 = N\pi(1-\pi) = 100(0.5)(1-0.5) = 25.$$

Therefore, $\sigma = 5$. For a σ of 5 and an N of 9, the standard deviation of the sampling distribution of the mean is $5/3 = 1.667$. Recall that the standard deviation of a sampling distribution is called the standard error.

To recap, we wish to know the probability of obtaining a sample mean of 51 or more when the sampling distribution of the mean has a mean of 50 and a standard deviation of 1.667. To compute this probability, we will make the assumption that the sampling distribution of the mean is normally distributed. We can then use a normal distribution calculator as shown in Figure 7.3.

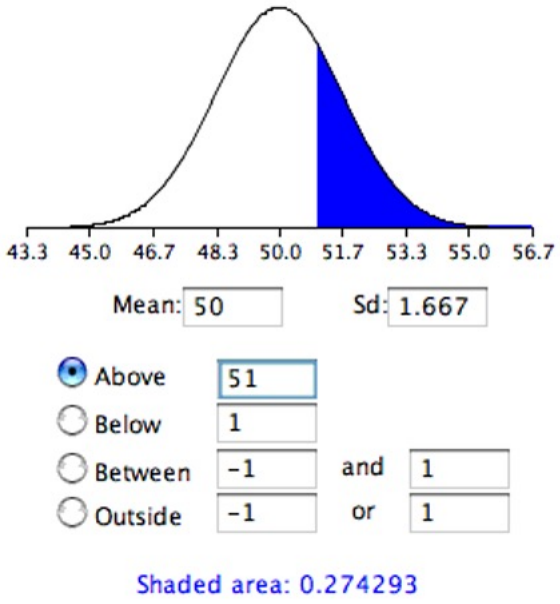


Figure 7.3: Probability of a sample mean being 51 or greater.

Notice that the mean is set to 50, the standard deviation to 1.667, and the area above 51 is requested and shown to be 0.274.

Therefore, the probability of obtaining a sample mean of 51 or larger is 0.274. Since a mean of 51 or higher is not unlikely under the assumption that the subliminal message has no effect, the effect is not significant and the null hypothesis is not rejected.

The test conducted above was a one-tailed test because it computed the probability of a sample mean being one or more points higher than the hypothesized mean of 50 and the area computed was the area above 51. To test the two-tailed hypothesis, you would compute the probability of a sample mean differing by one or more in either direction from the hypothesized mean of 50. You would do so by computing the probability of a mean being less than or equal to 49 or greater than or equal to 51.

The results from a normal distribution calculator are shown in Figure 7.4.

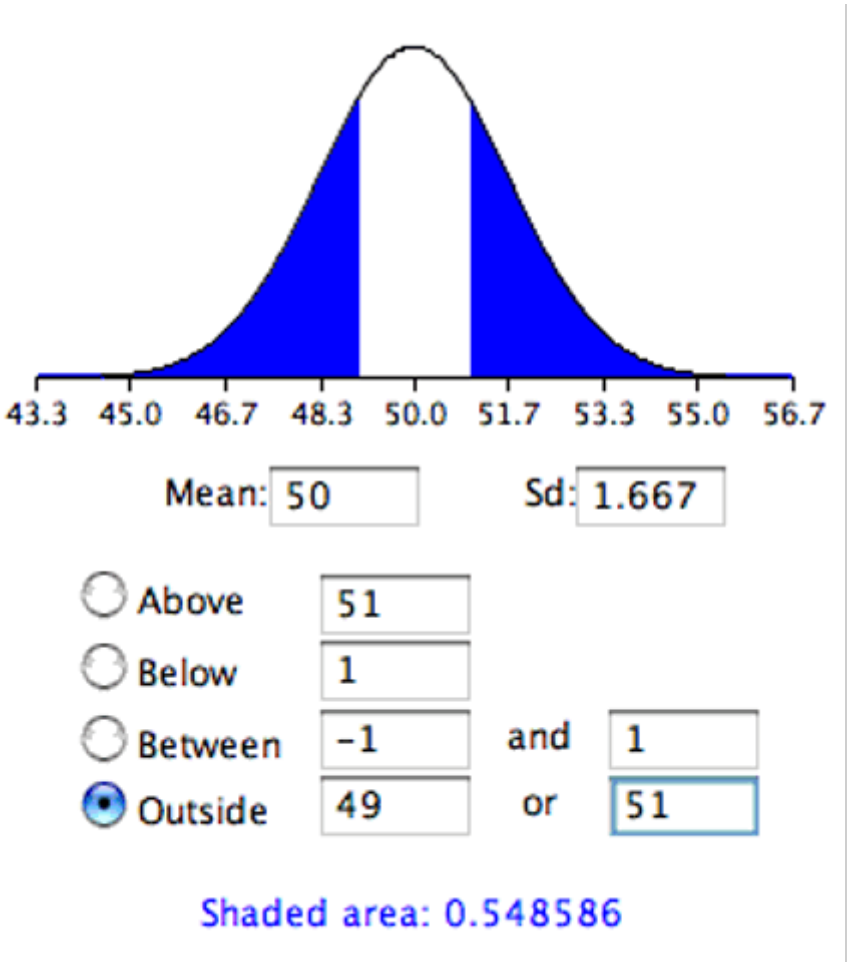


Figure 7.4: Probability of a sample mean being less than or equal to 49 or greater than or equal to 51.

As you can see, the probability is 0.548 which, as expected, is twice the probability of 0.274 shown in Figure 7.3.

Before normal calculators such as the one illustrated above were widely available, probability calculations were made based on the standard normal distribution. This was done by

computing Z based on the formula

$$Z = \frac{M - \mu}{\sigma_M}$$

where Z is the value on the standard normal distribution, M is the sample mean, μ is the hypothesized value of the mean, and σ_M is the standard error of the mean. For this example, Z = (51-50)/1.667 = 0.60. Use a normal calculator, with a mean of 0 and a standard deviation of 1, as shown below.

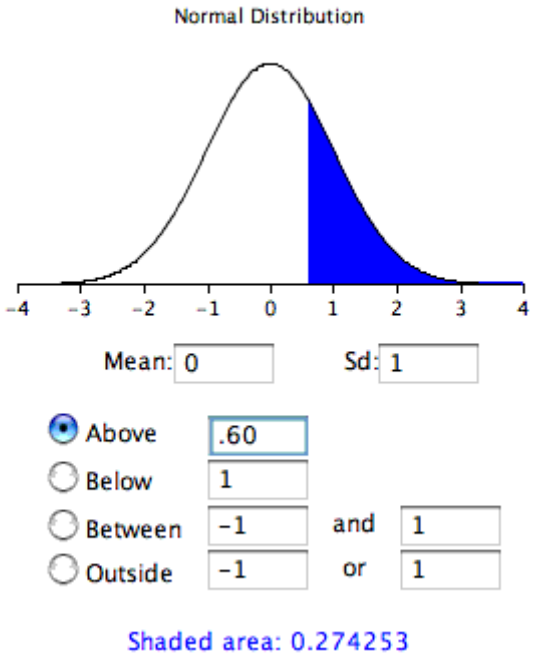


Figure 7.5: Calculation using the standardized normal distribution.

Notice that the probability (the shaded area) is the same as previously calculated (for the one-tailed test).

As noted, in real-world data analyses it is very rare that you would know σ and wish to estimate μ. Typically σ is not known and is estimated in a sample by s, and σ_M is estimated by s_M. For our next example, we will consider the data in the "ADHD Treatment" case study.[11] These data consist of the scores of 24 children with ADHD on a delay of gratification (DOG) task. Each child was tested under four dosage levels. Table 7.2 shows the data for the placebo (0 mg) and highest dosage level (0.6 mg) of methylphenidate. Of particular interest here is the column labeled "Diff" that shows the difference in performance between the 0.6 mg (D60) and the 0 mg (D0) conditions. These difference scores are positive for children who performed better in

the 0.6 mg condition than in the control condition and negative for those who scored better in the control condition. If methylphenidate has a positive effect, then the mean difference score in the population will be positive. The null hypothesis is that the mean difference score in the population is 0.

Table 7.2: DOG scores as a function of dosage.

D0	**D60**	**Diff**
57	62	5
27	49	22
32	30	-2
31	34	3
34	38	4
38	36	-2
71	77	6
33	51	18
34	45	11
53	42	-11
36	43	7
42	57	15
26	36	10
52	58	6
36	35	-1
55	60	5
36	33	-3
42	49	7
36	33	-3
54	59	5
34	35	1
29	37	8
33	45	12
33	29	-4

To test this null hypothesis, we compute t using a special case of the following formula:

$$\text{t} = \frac{\text{statistic-hypothesized value}}{\text{estimated standard error of the statistic}}$$

The special case of this formula applicable to testing a single mean is

$$\text{t} = \frac{M - \mu}{S_M}$$

where t is the value we compute for the significance test, M is the sample mean, μ is the hypothesized value of the population mean, and s_M is the estimated standard error of the mean. Notice the similarity of this formula to the formula for Z we saw before.

In the previous example, we assumed that the scores were normally distributed. In this case, it is the population of difference scores that we assume to be normally distributed.

The mean (M) of the N = 24 difference scores is 4.958, the hypothesized value of μ is 0, and the standard deviation (s) is 7.538. The estimate of the standard error of the mean is computed as:

Therefore, t = 4.96/1.54 = 3.22. The probability value for t depends on the degrees of freedom. The number of degrees of freedom is equal to N - 1 = 23. As shown below, a t distribution calculator finds that the probability of a t less than -3.22 or greater than 3.22 is only 0.0038. Therefore, if the drug had no effect, the probability of finding a difference between means as large or larger (in either direction) than the difference found is very low. Therefore the null hypothesis that the population mean difference score is zero can be rejected. The conclusion is that the population mean for the drug condition is higher than the population mean for the placebo condition.

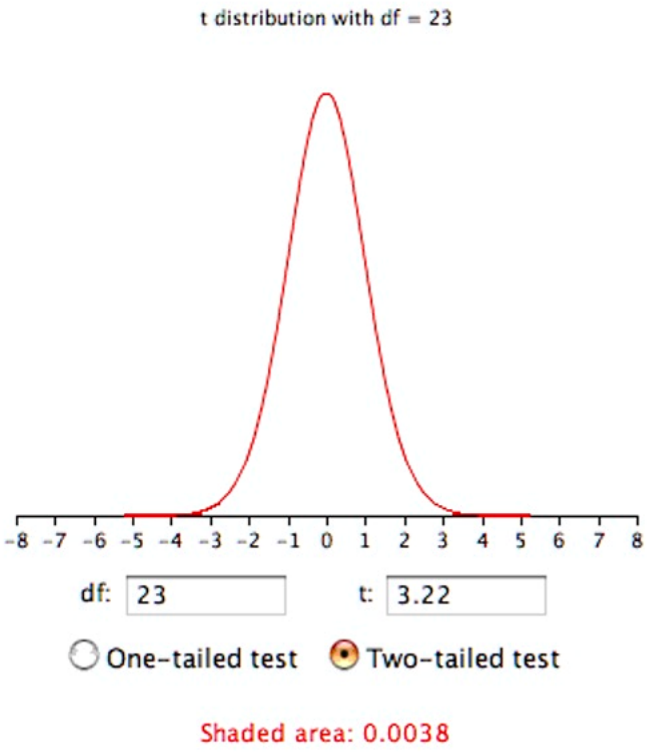


Figure 7.6: Calculation using the t distribution.

In order to conduct this hypothesis test, we made the following *assumptions*:

1. Each value is sampled independently from each other value.
2. The values are sampled from a normal distribution.

7.6 Type I and Type II Errors[12]

In the Physicians' Reactions case study,[13] the probability value associated with the significance test is 0.0057. Therefore, the null hypothesis was rejected, and it was concluded that physicians intend to spend less time with obese patients. Despite the low probability value, it is possible that the null hypothesis of no true difference between obese and average-weight patients is true and that the large difference between sample means occurred by chance. If this is the case, then the conclusion that physicians intend to spend less time with obese patients is in error. This type of error is called a Type I error. More generally, a **Type I error** occurs when a significance test results in the rejection of a true null hypothesis.

By one common convention, if the probability value is below 0.05, then the null hypothesis is rejected. Another convention, although slightly less common, is to reject the null hypothesis if the probability value is below 0.01. The threshold for rejecting the null hypothesis is called the α (alpha) level or simply α. It is also called the significance level. As discussed in the section on significance testing, it is better to interpret the probability value as an indication of the weight of evidence against the null hypothesis than as part of a decision rule for making a reject or do-not-reject decision. Therefore, keep in mind that rejecting the null hypothesis is not an all-or-nothing decision.

The Type I error rate is affected by the α level: the lower the α level, the lower the Type I error rate. It might seem that α is the probability of a Type I error. However, this is not correct. Instead, α is the probability of a Type I error given that the null hypothesis is true. If the null hypothesis is false, then it is impossible to make a Type I error.

The second type of error that can be made in significance testing is failing to reject a false null hypothesis. This kind of error is called a **Type II error**. Unlike a Type I error, a Type II error is not really an error. When a statistical test is not significant, it means that the data do not provide strong evidence that the null hypothesis is false. Lack of significance does not support the conclusion that the null hypothesis is true. Therefore, a researcher should not make the mistake of incorrectly concluding that the null hypothesis is true when a statistical test was not significant. Instead, the researcher should consider the test inconclusive. Contrast this with a Type I error in which the researcher erroneously concludes that the null hypothesis is false when, in fact, it is true.

A Type II error can only occur if the null hypothesis is false. If the null hypothesis is false, then the probability of a Type II error is called (beta). The probability of correctly rejecting a false null hypothesis equals 1- and is called *statistical power*.

[1] This section is adapted from David M. Lane. "Introduction." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/logic_of_hypothesis_testing/intro.html

[2] http://onlinestatbook.com/2/calculators/binomial_dist.html

[3] http://onlinestatbook.com/2/case_studies/weight.html

[4] Kaiser, H. F. (1960) Directional statistical decisions. *Psychological Review*, *67*, 160-167.

[5] This section is adapted from David M. Lane. "Steps in Hypothesis Testing." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/logic_of_hypothesis_testing/steps.html

[6] This section is adapted from David M. Lane. "One- and Two-Tailed Tests." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/logic_of_hypothesis_testing/tails.html

[7] http://onlinestatbook.com/2/case_studies/bond.html

[8] This section is adapted from David M. Lane. "Significance Testing." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/logic_of_hypothesis_testing/significance.html

[9] http://onlinestatbook.com/2/case_studies/weight.html

[10] This section is adapted from David M. Lane. "Testing a Single Mean." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/tests_of_means/single_mean.html

[11] http://onlinestatbook.com/2/case_studies/adhd.html

[12] This section is adapted from David M. Lane. "Type I and Type II Errors." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/logic_of_hypothesis_testing/errors.html

[13] http://onlinestatbook.com/2/case_studies/weight.html

8 Comparing Means (How a Qualitative Variable Relates to a Quantitative One)

8.1 Difference between Two Means[1]

It is much more common for a researcher to be interested in the difference between means than in the specific values of the means themselves. This section covers how to test for differences between means from two separate groups of subjects. Note that we already learned in the chapter on graphing how to visually depict a comparison of two means/medians using boxplots (Section 1.4.2).

We take as an example the data from the "Animal Research" case study, previously described when discussing confidence intervals (Section 4.3).[2] As a reminder, students rated (on a 7-point scale) whether they thought animal research is wrong. The sample sizes, means, and variances are shown separately for males and females in Table 8.1.

Table 8.1: Means and Variances in Animal Research study.

Group	n	Mean	Variance
Females	17	5.353	2.743
Males	17	3.882	2.985

As you can see, the females rated animal research as more wrong than did the males. This sample difference between the female mean of 5.35 and the male mean of 3.88 is 1.47. However, the gender difference in this particular sample is not very important. What is important is whether there is a difference in the *population* means.

In order to test whether there is a difference between population means, we are going to make three assumptions:

1. The two populations have the same variance. This assumption is called the assumption of homogeneity of variance.
2. The populations are normally distributed.

3. Each value is sampled independently from each other value. This assumption requires that each subject provide only one value. If a subject provides two scores, then the scores are not independent.

One could look at these assumptions in much more detail, but suffice it to say that small-to-moderate violations of assumptions 1 and 2 do not make much difference. It is important not to violate assumption 3.

We saw the following general formula for significance testing in the section on testing a single mean:

$$\text{t} = \frac{\text{statistic-hypothesized value}}{\text{estimated standard error of the statistic}}$$

In this case, our statistic is the difference between sample means and our hypothesized value is 0. The hypothesized value is the null hypothesis that the difference between population means is 0.

We continue to use the data from the "Animal Research" case study and will compute a significance test on the difference between the mean score of the females and the mean score of the males. For this calculation, we will make the three assumptions specified above.

The first step is to compute the statistic, which is simply the difference between means.

$$M_1 - M_2 = 5.3529 - 3.8824 = 1.4705$$

Since the hypothesized value is 0, we do not need to subtract it from the statistic.

The next step is to compute the estimate of the standard error of the statistic. In this case, the statistic is the difference between means, so the estimated standard error of the statistic is $(S_{M1} - _{M2})$. Recall from the relevant section in the chapter on sampling distributions that the formula for the standard error of the difference between means is:

$$\sigma_{M1-M2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1 2}} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}}$$

In order to estimate this quantity, we estimate ^{2} and use that estimate in place of ^{2}. Since we are assuming the two population variances are the same, we estimate this variance by averaging our two sample variances. Thus, our estimate of variance is computed using the following formula:

$$\text{MSE} = \frac{s_1^2 + s_2^2}{2}$$

where MSE is our estimate of σ^2. In this example,

$$\text{MSE} = (2.743 + 2.985)/2 = 2.864.$$

Since n (the number of scores in each group) is 17,

$$S_{M1-M2} = \sqrt{\frac{2MSE}{n}} = \sqrt{\frac{(2)(2.864)}{17}} = 0.5805$$

The next step is to compute t by plugging these values into the formula:

$$t = 1.4705/.5805 = 2.533.$$

Finally, we compute the probability of getting a t as large or larger than 2.533 or as small or smaller than -2.533. To do this, we need to know the degrees of freedom. The degrees of freedom is the number of independent estimates of variance on which MSE is based. This is equal to (n_1 - 1) + (n_2 - 1), where n_1 is the sample size of the first group and n_2 is the sample size of the second group. For this example, $n_1 = n_2 = 17$. When $n_1 = n_2$, it is conventional to use "n" to refer to the sample size of each group. Therefore, the degrees of freedom is 16 + 16 = 32.

Once we have the degrees of freedom, we can use a t distribution calculator[3] to find the probability. Figure 8.1 shows that the probability value for a two-tailed test is 0.0164. The two-tailed test is used when the null hypothesis can be rejected regardless of the direction of the effect. As shown in Figure 8.1, it is the probability of a t < -2.533 or a t > 2.533.

The results of a one-tailed test are shown in Figure 8.2. As you can see, the probability value of 0.0082 is half the value for the two-tailed test.

8.1.1 Formatting Data for Computer Analysis

Most computer programs that compute t tests require your data to be in a specific form. Consider the data in Table 8.2.

Table 8.2: Example data.

Group 1	**Group 2**
3	2
4	6
5	8

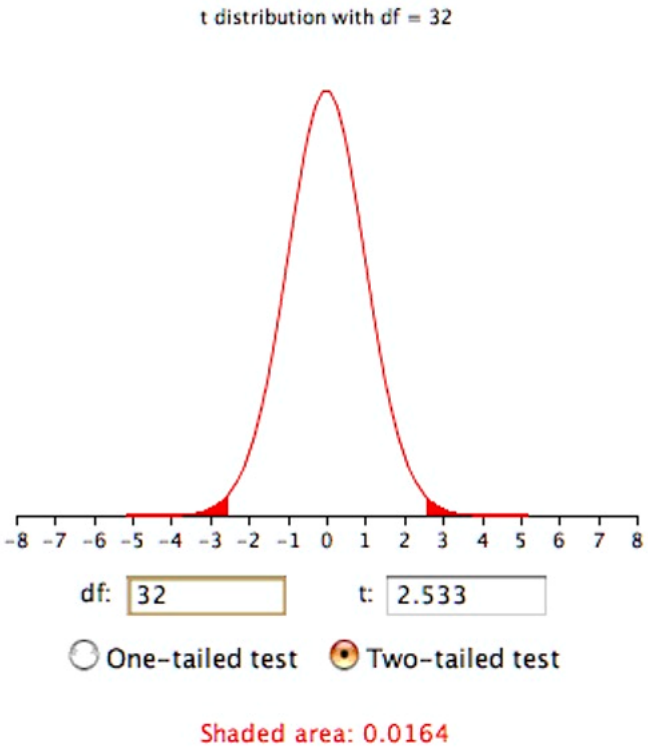


Figure 8.1: The two-tailed probability.

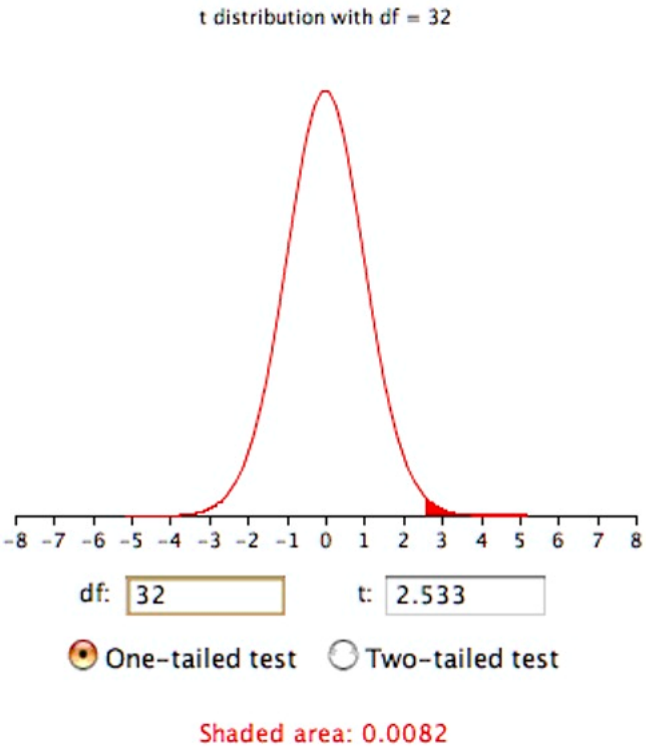


Figure 8.2: The one-tailed probability.

Here there are two groups, each with three observations. To format these data for a computer program, you normally have to use two variables: the first specifies the group the subject is in and the second is the score itself. The reformatted version of the data in Table 8.2 is shown in Table 8.3.

Table 8.3: Reformatted data.

G	Y
1	3
1	4
1	5
2	2
2	6
2	8

Using statistical software, we'd find that the t value is -0.718, the df = 4, and p = 0.512.

8.2 Pairwise Comparisons Among Multiple Means[4]

Many experiments are designed to compare more than two conditions. We will take as an example the case study "Smiles and Leniency."[5] In this study, the effect of different types of smiles on the leniency shown to a person was investigated. An obvious way to proceed would be to do a t test of the difference between each group mean and each of the other group means. This procedure would lead to the six comparisons shown in Table 8.4.

Table 8.4: Six Comparisons among Means.

felt vs. miserable		

The problem with this approach is that if you did this analysis, you would have six chances to make a Type I error. Therefore, if you were using the 0.05 significance level, the probability that you would make a Type I error on at least one of these comparisons is greater than 0.05. The more means that are compared, the more the Type I error rate is inflated. Figure 8.3 shows the number of possible comparisons between pairs of means (pairwise comparisons) as a function of the number of means. If there are only two means, then only one comparison can be made. If there are 12 means, then there are 66 possible comparisons.

Figure 8.4 shows the probability of a Type I error as a function of the number of means. As you can see, if you have an experiment with 12 means, the probability is about 0.70 that at least one of the 66 comparisons among means would be significant even if all 12 population means were the same.

The Type I error rate can be controlled using a test called the Tukey Honestly Significant Difference test or Tukey HSD for short. The Tukey HSD test is one example of a multiple comparison test, but several alternatives are frequently used, such as the Bonferroni correction. Regardless of the exact method used for a multiple comparison test, the interpretation of results is similar. The Tukey HSD is based on a variation of the t distribution that takes into account the number of means being compared. This distribution is called the studentized range distribution.

Normally, statistical software will make all the necessary calculations for you in the background. But to illustrate what sorts of calculations the software is relying on, let's return to the leniency study to see how to compute the Tukey HSD test. You will see that the computations are very similar to those of an independent-groups t test. The steps are outlined below:

1. Compute the means and variances of each group. They are shown below.

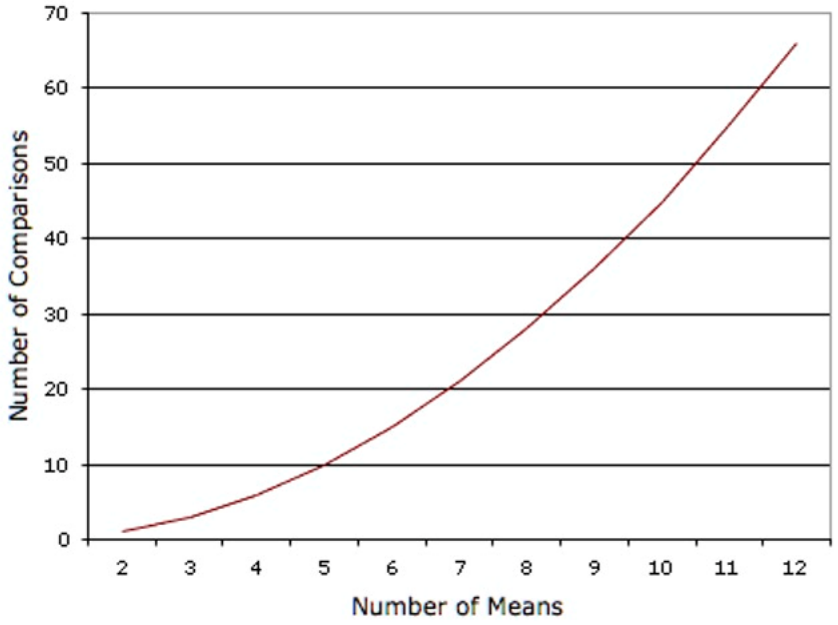


Figure 8.3: Number of pairwise comparisons as a function of the number of means.

Condition	Mean	Variance
False	5.37	3.34
Felt	4.91	2.83
Miserable	4.91	2.11
Neutral	4.12	2.32

2. Compute MSE, which is simply the mean of the variances. It is equal to 2.65.

3. Compute

$$Q = \frac{M_i - M_j}{\sqrt{\frac{MSE}{n}}}$$

for each pair of means, where M_i is one mean, M_j is the other mean, and n is the number of scores in each group. For these data, there are 34 observations per group. The value in the denominator is 0.279.

4. Compute p for each comparison using a Studentized Range Calculator.[6] The degrees of freedom is equal to the total number of observations minus the number of means. For this experiment, df = 136 - 4 = 132.

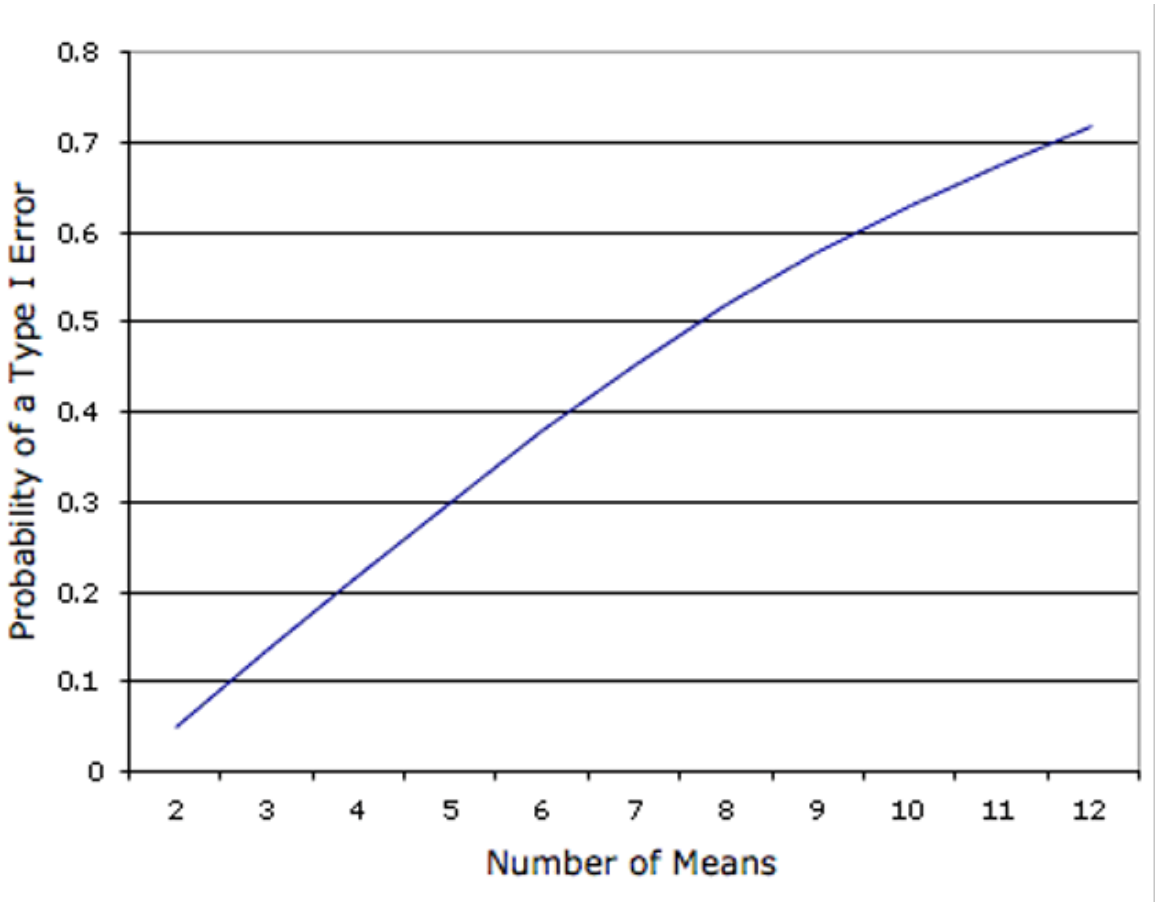


Figure 8.4: Probability of a Type I error as a function of the number of means.

The tests for these data are shown in Table 8.6.

Table 8.6: Six Pairwise Comparisons.

Comparison	$\mathbf{M_i}$-$\mathbf{M_j}$	**Q**	**p**
False - Felt	0.46	1.65	0.649
False - Miserable	0.46	1.65	0.649
False - Neutral	1.25	4.48	0.010
Felt - Miserable	0.00	0.00	1.000
Felt - Neutral	0.79	2.83	0.193
Miserable - Neutral	0.79	2.83	0.193

The only significant comparison is between the false smile and the neutral smile.

It is not unusual to obtain results that on the surface appear paradoxical. For example, these results appear to indicate that (a) the false smile is the same as the miserable smile, (b) the miserable smile is the same as the neutral control, and (c) the false smile is different from the neutral control. This apparent contradiction is avoided if you are careful not to accept the null hypothesis when you fail to reject it. The finding that the false smile is not significantly different from the miserable smile does not mean that they are really the same. Rather it means that there is not convincing evidence that they are different. Similarly, the non-significant difference between the miserable smile and the control does not mean that they are the same. The proper conclusion is that the false smile is higher than the control and that the miserable smile is either (a) equal to the false smile, (b) equal to the control, or (c) somewhere in-between.

The assumptions of the Tukey test are essentially the same as for an independent-groups t test: normality, homogeneity of variance, and independent observations. The test is quite robust to violations of normality. Violating homogeneity of variance can be more problematical than in the two-sample case since the MSE is based on data from all groups. The assumption of independence of observations is important and should not be violated.

8.2.1 Computer Analysis

For most computer programs, you should format your data the same way you do for an independent-groups t test. The only difference is that if you have, say, four groups, you would code each group as 1, 2, 3, or 4 rather than just 1 or 2.

8.2.2 Tukey's Test Need Not be a Follow-Up to ANOVA

Some textbooks introduce the Tukey test only as a follow-up to an analysis of variance. There is no logical or statistical reason why you should not use the Tukey test even if you do not

compute an ANOVA (or even know what one is). If you or your instructor do not wish to take our word for this, see the excellent article on this and other issues in statistical analysis by Leland Wilkinson and the APA Board of Scientific Affairs' Task Force on Statistical Inference, published in the *American Psychologist*, August 1999, Vol. *54*, No. 8, 594–604.

8.3 Analysis of Variance (ANOVA)[7]

Analysis of Variance (ANOVA) is a statistical method used to test differences between two or more means. It may seem odd that the technique is called "Analysis of Variance" rather than "Analysis of Means." As you will see, the name is appropriate because inferences about means are made by analyzing variance.

ANOVA is used to test general rather than specific differences among means. This can be seen best by example. In the case study "Smiles and Leniency,"[8] the effect of different types of smiles on the leniency shown to a person was investigated. Four different types of smiles (neutral, false, felt, miserable) were investigated. In the prior section, we learned how to test differences among means. The results from the Tukey HSD test are shown in Table 8.7.

Table 8.7: Six Pairwise Comparisons.

Comparison	**M_i-M_j**	**Q**	**p**
False - Felt	0.46	1.65	0.649
False - Miserable	0.46	1.65	0.649
False - Neutral	1.25	4.48	0.010
Felt - Miserable	0.00	0.00	1.000
Felt - Neutral	0.79	2.83	0.193
Miserable - Neutral	0.79	2.83	0.193

Notice that the only significant difference is between the False and Neutral conditions.

ANOVA tests the non-specific null hypothesis that all four population means are equal. That is,

$$\mu_{false} = \mu_{felt} = \mu_{miserable} = \mu_{neutral}$$

This non-specific null hypothesis is sometimes called the omnibus null hypothesis. When the omnibus null hypothesis is rejected, the conclusion is that at least one population mean is different from at least one other mean. However, since the ANOVA does not reveal which means are different from which, it offers less specific information than the Tukey HSD test. The Tukey HSD is therefore preferable to ANOVA in this situation. Some textbooks introduce

the Tukey test only as a follow-up to an ANOVA. However, there is no logical or statistical reason why you should not use the Tukey test even if you do not compute an ANOVA.

You might be wondering why you should learn about ANOVA when the Tukey test is better. One reason is that there are complex types of analyses that can be done with ANOVA and not with the Tukey test. A second is that ANOVA is by far the most commonly-used technique for comparing means, and it is important to understand ANOVA in order to understand research reports.

---

[1] This section is adapted from David M. Lane. "Difference between Two Means (Independent Groups)." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/tests_of_means/difference_means.html

[2] http://onlinestatbook.com/2/case_studies/animal_research.html

[3] http://onlinestatbook.com/2/calculators/t_dist.html

[4] This section is adapted from David M. Lane. "All Pairwise Comparisons Among Means." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/tests_of_means/pairwise.html

[5] http://onlinestatbook.com/2/case_studies/leniency.html

[6] http://onlinestatbook.com/2/calculators/studentized_range_dist.html

[7] This section is adapted from David M. Lane. "Introduction." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/analysis_of_variance/intro.html

[8] http://onlinestatbook.com/2/case_studies/leniency.html

9 Comparing Groups (How Two Qualitative Variables Relate to One Another)

We previously learned how to graphically depict the relationship between two qualitative variables (Section 1.3.3.1). To make a comparison that includes a significance test, we will need to use the chi square distribution, together with a contingency table.

9.1 Chi Square Distribution[1]

A standard normal deviate is a random sample from the standard normal distribution. The Chi Square distribution is the distribution of the sum of squared standard normal deviates. The degrees of freedom of the distribution is equal to the number of standard normal deviates being summed. Therefore, Chi Square with one degree of freedom, written as $^2(1)$, is simply the distribution of a single normal deviate squared. The area of a Chi Square distribution below 4 is the same as the area of a standard normal distribution below 2, since 4 is 2^2.

Consider the following problem: you sample two scores from a standard normal distribution, square each score, and sum the squares. What is the probability that the sum of these two squares will be six or higher? Since two scores are sampled, the answer can be found using the Chi Square distribution with two degrees of freedom. A Chi Square calculator can be used to find that the probability of a Chi Square (with 2 df) being six or higher is 0.050.

The mean of a Chi Square distribution is its degrees of freedom. Chi Square distributions are positively skewed, with the degree of skew decreasing with increasing degrees of freedom. As the degrees of freedom increases, the Chi Square distribution approaches a normal distribution. Figure 9.1 shows density functions for three Chi Square distributions. Notice how the skew decreases as the degrees of freedom increases.

The Chi Square distribution is very important because many test statistics are approximately distributed as Chi Square. Two of the more common tests using the Chi Square distribution are tests of deviations of differences between theoretically expected and observed frequencies (one-way tables) and the relationship between categorical variables (contingency tables). Numerous other tests beyond the scope of this work are based on the Chi Square distribution.

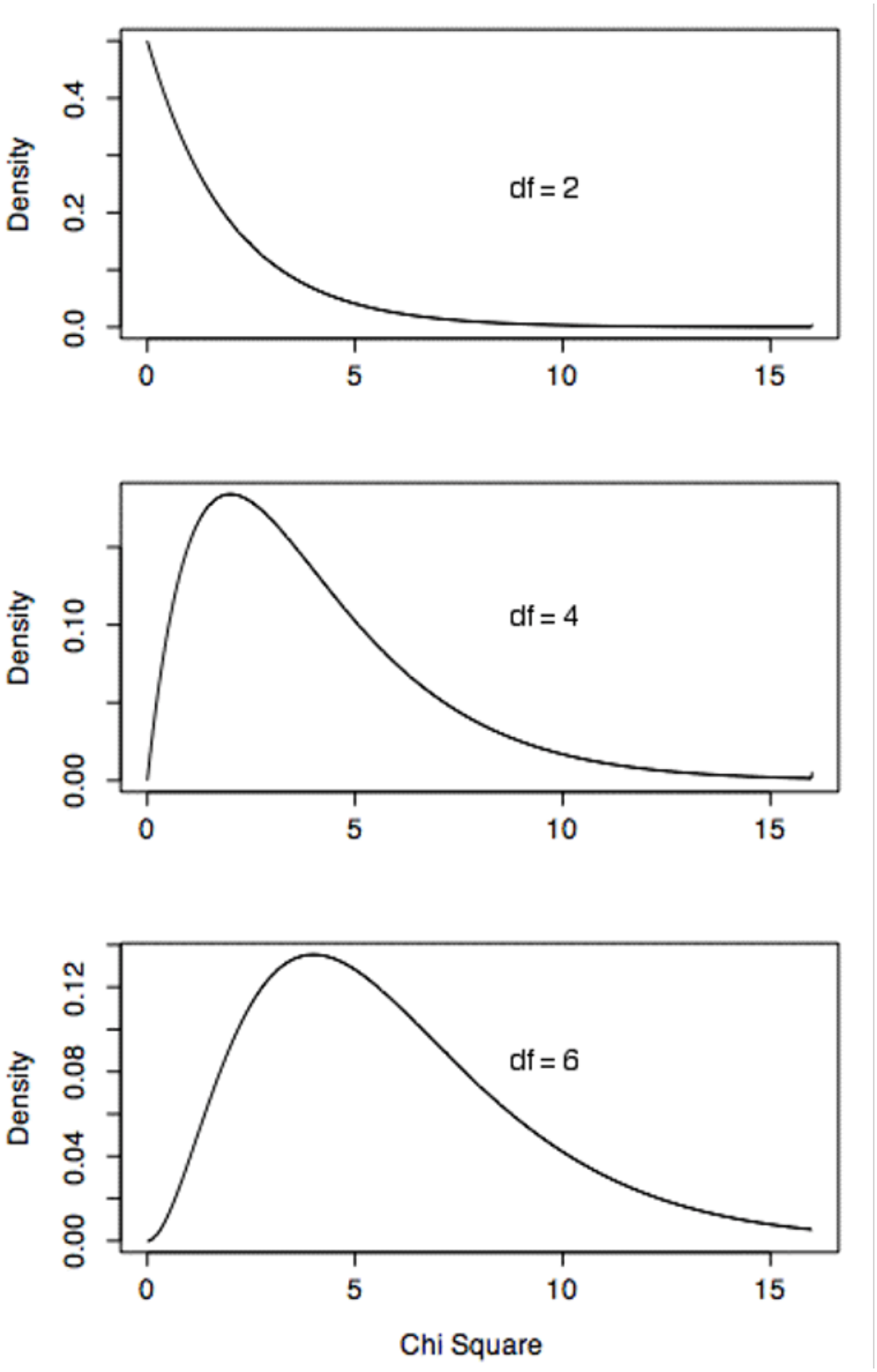


Figure 9.1: Chi Square distributions with 2, 4, and 6 degrees of freedom.

9.2 One-Way Tables[2]

The Chi Square distribution can be used to test whether observed data differ *significantly* from theoretical expectations. For example, for a fair six-sided die, the probability of any given outcome on a single roll would be 1/6. The data in Table 9.1 were obtained by rolling a six-sided die 36 times. However, as can be seen in Table 9.1, some outcomes occurred more frequently than others. For example, a "3" came up nine times, whereas a "4" came up only two times. Are these data consistent with the hypothesis that the die is a fair die? Naturally, we do not expect the sample frequencies of the six possible outcomes to be the same since chance differences will occur. So, the finding that the frequencies differ does not mean that the die is not fair. One way to test whether the die is fair is to conduct a significance (hypothesis) test. The null hypothesis is that the die is fair. This hypothesis is tested by computing the probability of obtaining frequencies as discrepant or more discrepant from a uniform distribution of frequencies as obtained in the sample. If this probability is sufficiently low, then the null hypothesis that the die is fair can be rejected.

Table 9.1: Outcome Frequencies from a Six-Sided Die.

Outcome	**Frequency**
1	8
2	5
3	9
4	2
5	7
6	5

The first step in conducting the significance test is to compute the expected frequency for each outcome given that the null hypothesis is true. For example, the expected frequency of a "1" is 6 since the probability of a "1" coming up is 1/6 and there were a total of 36 rolls of the die.

Expected frequency = $(1/6)(36) = 6$

Note that the expected frequencies are expected only in a theoretical sense. We do not really "expect" the observed frequencies to match the "expected frequencies" exactly.

The calculation continues as follows. Letting E be the expected frequency of an outcome and O be the observed frequency of that outcome, compute

$$\frac{(E-O)^2}{E}$$

for each outcome. Table 9.2 shows these calculations.

Table 9.2: Outcome Frequencies from a Six-Sided Die.

Outcome	E	O	$\frac{(E-O)^2}{E}$
1	6	8	0.667
2	6	5	0.167
3	6	9	1.500
4	6	2	2.667
5	6	7	0.167
6	6	5	0.167

Next we add up all the values in Column 4 of Table 9.2.

$$\Sigma \frac{(E-O)^2}{E} = 5.333$$

This sampling distribution of

$$\Sigma \frac{(E-O)^2}{E}$$

is approximately distributed as Chi Square with k-1 degrees of freedom, where k is the number of categories. Therefore, for this problem the test statistic is

$$X_5^2 = 5.333$$

which means the value of Chi Square with 5 degrees of freedom is 5.333.

From a Chi Square calculator[3] it can be determined that the probability of a Chi Square of 5.333 or larger is 0.377. Therefore, the null hypothesis that the die is fair cannot be rejected.

This Chi Square test can also be used to test other deviations between expected and observed frequencies. The following example shows a test of whether the variable "University GPA" in the SAT and College GPA case study is normally distributed.

The first column in Table 9.3 shows the normal distribution divided into five ranges. The second column shows the proportions of a normal distribution falling in the ranges specified in the first column. The expected frequencies (E) are calculated by multiplying the number of scores (105) by the proportion. The final column shows the observed number of scores in each range. It is clear that the observed frequencies vary greatly from the expected frequencies. Note that if the distribution were normal, then there would have been only about 35 scores between 0 and 1, whereas 60 were observed.

Table 9.3: Expected and Observed Scores for 105 University GPA Scores.

Range	**Proportion**	**E**	**O**
Above 1	0.159	16.695	9
0 to 1	0.341	35.805	60
-1 to 0	0.341	35.805	17
Below -1	0.159	16.695	19

The test of whether the observed scores deviate significantly from the expected scores is computed using the familiar calculation.

$$X_3^2 = \Sigma \frac{(E-O)^2}{E} = 30.09$$

The subscript "3" means there are three degrees of freedom. As before, the degrees of freedom is the number of outcomes minus 1, which is 4 - 1 = 3 in this example. A Chi Square distribution calculator shows that $p < 0.001$ for this Chi Square. Therefore, the null hypothesis that the scores are normally distributed can be rejected.

9.3 Contingency Tables[4]

This section shows how to use Chi Square to test the relationship between nominal variables for significance. For example, Table 9.4 shows the data from the Mediterranean Diet and Health case study.[5]

Table 9.4: Frequencies for Diet and Health Study.

		Outcome			
Diet	Cancers	Fatal Heart Disease	Non-Fatal Heart Disease	Healthy	Total
AHA	15	24	25	239	303
Mediterranean	7	14	8	273	302
Total	22	38	33	512	605

The question is whether there is a *significant relationship* between diet and outcome. Again, software can calculate a p-value for us in order to test for significance. But if we are wondering what's going on under the hood, the first step is to compute the expected frequency for each

cell based on the assumption that there is no relationship. These expected frequencies are computed from the totals as follows. We begin by computing the expected frequency for the AHA Diet/Cancers combination. Note that 22/605 subjects developed cancer. The proportion who developed cancer is therefore 0.0364. If there were no relationship between diet and outcome, then we would expect 0.0364 of those on the AHA diet to develop cancer. Since 303 subjects were on the AHA diet, we would expect (0.0364)(303) = 11.02 cancers on the AHA diet. Similarly, we would expect (0.0364)(302) = 10.98 cancers on the Mediterranean diet. In general, the expected frequency for a cell in the ith row and the jth column is equal to

$$E_{ij} = \frac{T_i T_j}{T}$$

where E_{ij} is the expected frequency for cell i,j, Ti is the total for the ith row, Tj is the total for the jth column, and T is the total number of observations. For the AHA Diet/Cancers cell, i = 1, j = 1, Ti = 303, Tj = 22, and T = 605. Table 9.5 shows the expected frequencies (in parenthesis) for each cell in the experiment.

Table 9.5: Observed and Expected Frequencies for Diet and Health Study.

	Outcome			
Diet	Cancers	Fatal Heart Disease	Non-Fatal Heart Disease	Healthy
AHA	15 (11.02)	24 (19.03)	25 (16.53)	239 (256.42)
Mediterranean	7 (10.98)	14 (18.97)	8 (16.47)	273 (255.58)
Total	22	38	33	512

The significance test is conducted by computing Chi Square as follows.

$$X_3^2 = \Sigma \frac{(E-O)^2}{E} = -16.55$$

The degrees of freedom is equal to (r-1)(c-1), where r is the number of rows and c is the number of columns. For this example, the degrees of freedom is (2-1)(4-1) = 3. The Chi Square calculator[6] can be used to determine that the probability value for a Chi Square of 16.55 with three degrees of freedom is equal to 0.0009. Therefore, the null hypothesis of no relationship between diet and outcome can be rejected.

A key assumption of this Chi Square test is that each subject contributes data to only one cell. Therefore, the sum of all cell frequencies in the table must be the same as the number of subjects in the experiment. Consider an experiment in which each of 16 subjects attempted two anagram problems. The data are shown in Table 9.6.

Table 9.6: Anagram Problem Data.

	Anagram 1	**Anagram 2**
Solved	10	4
Did not Solve	6	12

It would not be valid to use the Chi Square test on these data since each subject contributed data to two cells: one cell based on their performance on Anagram 1 and one cell based on their performance on Anagram 2. The total of the cell frequencies in the table is 32, but the total number of subjects is only 16.

---

[1] This section is adapted from David M. Lane. "Chi Square Distribution." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/chi_square/distribution.html

[2] This section is adapted from David M. Lane. "One-Way Tables (Testing Goodness of Fit)." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/chi_square/one-way.html

[3] http://onlinestatbook.com/2/calculators/chi_square_prob.html

[4] This section is adapted from David M. Lane. "Contingency Tables." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/chi_square/contingency.html

[5] http://onlinestatbook.com/2/case_studies/diet.html

[6] http://onlinestatbook.com/2/calculators/chi_square_prob.html

10 Causality

10.1 Causation[1]

The concept of causation is a complex one in the philosophy of science.[2] Since a full coverage of this topic is well beyond the scope of this text, we focus on two specific topics: (1) the establishment of causation in experiments and (2) the establishment of causation in non-experimental designs.

10.1.1 Establishing Causation in Experiments

Consider a simple experiment in which subjects are sampled randomly from a population and then assigned randomly to either the experimental group or the control group. Assume the condition means on the dependent variable differed. Does this mean the treatment caused the difference?

To make this discussion more concrete, assume that the experimental group received a drug for insomnia, the control group received a placebo, and the dependent variable was the number of minutes the subject slept that night. An obvious obstacle to inferring causality is that there are many unmeasured variables that affect how many hours someone sleeps. Among them are how much stress the person is under, physiological and genetic factors, how much caffeine they consumed, how much sleep they got the night before, etc. Perhaps differences between the groups on these factors are responsible for the difference in the number of minutes slept.

At first blush it might seem that the random assignment eliminates differences in unmeasured variables. However, this is not the case. Random assignment ensures that differences on unmeasured variables are chance differences. It does not ensure that there are no differences. Perhaps, by chance, many subjects in the control group were under high stress and this stress made it more difficult to fall asleep. The fact that the greater stress in the control group was due to chance does not mean it could not be responsible for the difference between the control and the experimental groups. In other words, the observed difference in "minutes slept" could have been due to a chance difference between the control group and the experimental group rather than due to the drug's effect.

[1]This section is adapted from David M. Lane. "Causation." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/research_design/causation.html

[2]See http://plato.stanford.edu/search/searcher.py?query=causation

This problem seems intractable since, by definition, it is impossible to measure an "unmeasured variable" just as it is impossible to measure and control all variables that affect the dependent variable. However, although it is impossible to assess the effect of any single unmeasured variable, it is possible to assess the combined effects of all unmeasured variables. Since everyone in a given condition is treated the same in the experiment, differences in their scores on the dependent variable must be due to the unmeasured variables. Therefore, a measure of the differences among the subjects within a condition is a measure of the sum total of the effects of the unmeasured variables. The most common measure of differences is the variance. By using the within-condition variance to assess the effects of unmeasured variables, statistical methods determine the probability that these unmeasured variables could produce a difference between conditions as large or larger than the differenc[3]e obtained in the experiment. If that probability is low, then it is inferred (that's why they call it inferential statistics) that the treatment had an effect and that the differences are not entirely due to chance. Of course, there is always some nonzero probability that the difference occurred by chance so total certainty is not a possibility.

10.1.2 Causation in Non-Experimental Designs

It is almost a cliché that correlation does not mean causation. The main fallacy in inferring causation from correlation is called the third variable problem and means that a third variable is responsible for the correlation between two other variables. An excellent example used by Li (1975) to illustrate this point is the positive correlation in Taiwan in the 1970's between the use of contraception and the number of electric appliances in one's house. Of course, using contraception does not induce you to buy electrical appliances or vice versa. Instead, the third variable of education level affects both.

Does the possibility of a third-variable problem make it impossible to draw causal inferences without doing an experiment? One approach is to simply assume that you do not have a third-variable problem. This approach, although common, is not very satisfactory. However, be aware that the assumption of no third-variable problem may be hidden behind a complex causal model that contains sophisticated and elegant mathematics.

A better though, admittedly more difficult approach, is to find converging evidence. This was the approach taken to conclude that smoking causes cancer. The analysis included converging evidence from retrospective studies, prospective studies, lab studies with animals, and theoretical understandings of cancer causes.

A second problem is determining the direction of causality. A correlation between two variables does not indicate which variable is causing which. For example, Reinhart and Rogoff (2010)[4] found a strong correlation between public debt and GDP growth. Although some have

[3]Li, C. (1975) *Path analysis: A primer.* Boxwood Press, Pacific Grove, CA.

[4]Reinhart, C. M. and Rogoff, K. S. (2010). Growth in a Time of Debt. Working Paper 15639, National Bureau of Economic Research, http://www.nber.org/papers/w15639

argued that public debt slows growth, most evidence supports the alternative that slow growth increases public debt.[5]

10.2 Experimental Designs[6]

There are many ways an experiment can be designed. For example, subjects can all be tested under each of the treatment conditions or a different group of subjects can be used for each treatment. An experiment might have just one independent variable or it might have several. This section describes basic experimental designs and their advantages and disadvantages.

10.2.1 Between-Subjects Designs

In a **between-subjects** design, the various experimental treatments are given to different groups of subjects. For example, in the "Teacher Ratings"[7] case study, subjects were randomly divided into two groups. Subjects were all told they were going to see a video of an instructor's lecture after which they would rate the quality of the lecture. The groups differed in that the subjects in one group were told that prior teaching evaluations indicated that the instructor was charismatic whereas subjects in the other group were told that the evaluations indicated the instructor was punitive. In this experiment, the independent variable is "Condition" and has two levels (charismatic teacher and punitive teacher). It is a between-subjects variable because different subjects were used for the two levels of the independent variable: subjects were in either the "charismatic teacher" or the "punitive teacher" condition. Thus the comparison of the charismatic-teacher condition with the punitive-teacher condition is a comparison between the subjects in one condition with the subjects in the other condition.

The two conditions were treated exactly the same except for the instructions they received. Therefore, it would appear that any difference between conditions should be attributed to the treatments themselves. However, this ignores the possibility of chance differences between the groups. That is, by chance, the raters in one condition might have, on average, been more lenient than the raters in the other condition. Randomly assigning subjects to treatments ensures that all differences between conditions are chance differences; it does not ensure there will be no differences. The key question, then, is how to distinguish real differences from chance differences. The field of inferential statistics answers just this question. Analyzing the data from this experiment reveals that the ratings in the charismatic-teacher condition were higher than those in the punitive-teacher condition. Using inferential statistics, it can be calculated that the probability of finding a difference as large or larger than the one obtained

[5] For a video on causality featuring evidence that smoking causes cancer, see http://www.learner.org/resources/series65.html

[6] This section is adapted from David M. Lane. "Experimental Designs." *Online Statistics Education: A Multimedia Course of Study.* http://onlinestatbook.com/2/research_design/designs.html

[7] http://onlinestatbook.com/2/case_studies/ratings.html

if the treatment had no effect is only 0.018. Therefore it seems likely that the treatment had an effect and it is not the case that all differences were chance differences.

Independent variables often have several levels. For example, in the "Smiles and Leniency" case study the independent variable is "type of smile" and there are four levels of this independent variable: (1) false smile, (2) felt smile, (3) miserable smile, and (4) a neutral control. Keep in mind that although there are four levels, there is only one independent variable. Designs with more than one independent variable are considered next.

10.2.2 Multi-Factor Between-Subject Designs

In the "Bias Against Associates of the Obese"[8] experiment, the qualifications of potential job applicants were judged. Each applicant was accompanied by an associate. The experiment had two independent variables: the weight of the associate (obese or average) and the applicant's relationship to the associate (girl friend or acquaintance). This design can be described as an Associate's Weight (2) x Associate's Relationship (2) factorial design. The numbers in parentheses represent the number of levels of the independent variable. The design was a factorial design because all four combinations of associate's weight and associate's relationship were included. The dependent variable was a rating of the applicant's qualifications (on a 9-point scale).

If two separate experiments had been conducted, one to test the effect of Associate's Weight and one to test the effect of Associate's Relationship then there would be no way to assess whether the effect of Associate's Weight depended on the Associate's Relationship. One might imagine that the Associate's Weight would have a larger effect if the associate were a girl friend rather than merely an acquaintance. A factorial design allows this question to be addressed. When the effect of one variable does differ depending on the level of the other variable then it is said that there is an interaction between the variables.

Factorial designs can have three or more independent variables. In order to be a between-subjects design there must be a separate group of subjects for each combination of the levels of the independent variables.

10.2.3 Within-Subjects Designs

A **within-subjects** design differs from a between-subjects design in that the same subjects perform at all levels of the independent variable. For example consider the "ADHD Treatment"[9] case study. In this experiment, subjects diagnosed as having attention deficit disorder were each tested on a delay of gratification task after receiving methylphenidate (MPH). All

[8] http://onlinestatbook.com/2/case_studies/obesity_relation.html
[9] http://onlinestatbook.com/2/case_studies/adhd.html

subjects were tested four times, once after receiving one of the four doses. Since each subject was tested under *each* of the four levels of the independent variable "dose," the design is a within-subjects design and dose is a within-subjects variable. Within-subjects designs are sometimes called repeated-measures designs.

10.2.4 Advantage of Within-Subjects Designs

An advantage of within-subjects designs is that individual differences in subjects' overall levels of performance are controlled. This is important because subjects invariably will differ greatly from one another. In an experiment on problem solving, some subjects will be better than others regardless of the condition they are in. Similarly, in a study of blood pressure some subjects will have higher blood pressure than others regardless of the condition. Within-subjects designs control these individual differences by comparing the scores of a subject in one condition to the scores of the same subject in other conditions. In this sense each subject serves as his or her own control. This typically gives within-subjects designs considerably more power (ability to find precise estimates) than between-subjects designs. That is, this makes within-subjects designs more able to detect an effect of the independent variable than are between-subjects designs.

Within-subjects designs are often called "repeated-measures" designs since repeated measurements are taken for each subject. Similarly, a within-subject variable can be called a repeated-measures factor.

10.2.5 Complex Designs

Designs can contain combinations of between-subject and within-subject variables. For example, the "Weapons and Aggression"[10] case study has one between-subject variable (gender) and two within-subject variables (the type of priming word and the type of word to be responded to).

[10] http://onlinestatbook.com/2/case_studies/guns.html

11 Models and Uncertainty

Before I leave my house each morning, I need to decide whether to take an umbrella. So I check my phone to see whether it's supposed to rain. Instead of giving me a direct yes or no answer, the weather tells me the percent chance of rain for the day.

Why does the weather app give me a percentage? Because there's uncertainty. Science has done a lot to help us understand the weather. And as our understanding of the weather improves, our predictions get better. But we still can't predict rain perfectly.

Facing uncertainty is a common problem when we're looking at data. Whether we're trying to explain the weather, human behavior, or even plant growth, we can't make perfect predictions because there are things we can't fully explain with our current scientific knowledge.

In statistics, we have several tools that allow us to acknowledge uncertainty. This enables us to build models like the ones powering my weather app—models that give us a prediction that includes a description of how uncertain we are. Some days we are 100% sure it will rain, other days only 60%.

In order to build these models that acknowledge uncertainty, we need a way to talk about what we do know and what we don't know. Let me give a very simple example of a model that accounts for uncertainty:

$$happiness = 3.0 + 2.3 \times income + \varepsilon \tag{11.1}$$

This model attempts to explain one's level of happiness based on their income. You might notice that it looks very similar to the regression equations we saw in Chapter 3. That's because regression is one of the main tools used to estimate a model that includes uncertainty.

What does this model mean in practical terms? Well, there are no obvious units we can use to quantify the amount of happiness someone experiences, so the exact values of the numbers we see are not particularly meaningful. But the fact that there's a positive number (2.3) that is being multiplied by income implies that as income gets bigger, happiness gets larger.

The key part of this equation that I want to focus on is the little Greek letter at the end of the equation: ε. This letter is called "epsilon," and it is often used to represent what we call an **error term** (also sometimes called a **disturbance term**). The error term (ε) represents everything else besides income that affects happiness. By including an error term, we are acknowledging that we can't perfectly predict one's level of happiness based on their income.

We think that knowing one's income will help us predict their happiness, but we know there are other factors we won't be able to measure or identify that will also affect happiness. Thus, if all we know about someone is their income, we will have uncertainty about their exact level of happiness. By including an error term (ε) in the model, we make clear that we only claim to have a partial understanding of happiness, not a complete one.

Think for a moment about how few topics we could study if we didn't have the freedom to build models that include uncertainty. We'd only be able to build a model of a dependent variable after we had identified (and measured) *all* of the factors that affect that variable! We wouldn't be able to build a model of rain since we don't know all of the factors that affect the rain. We couldn't build a model of voting behavior since we don't know everything that affects how someone will vote. By including an error term in our model, we can build models even when our understanding of something is incomplete.

The first part of our model that appears on the right side of the equation ($3.0 + 2.3 \times income$) is sometimes described as the *systematic* part of our model. It's what we would use to build a prediction of happiness if all we know about some is their income level. Suppose, for example, that someone has an income of 4 units (perhaps income is measured in tens of thousands of dollars of annual income, so a salary of $40,000 is coded as a 4). According to our model, that person's happiness would be:

$$happiness = 3.0 + 2.3 \times (4) + \varepsilon$$

$$happiness = 12.2 + \varepsilon$$

We, therefore, predict that someone with an income of 4 will have a happiness of 12.2, but we also acknowledge that their actual happiness will likely be a bit different from our prediction since our model indicates that their actual happiness will equal 12.2 plus the value of the error term (ε).

The error term describes something unknown, so we can't measure it or directly observe it. But what we can do is talk about its characteristics using concepts from probability theory. Specifically, we're going to describe the value of the error term as being randomly selected. You may have dealt with randomness in math classes before using examples such as coin flips, die rolls, or drawing cards from a 52-card deck. Just as the likelihood of different outcomes from parlor games can be described using probability, we're going to use probability to describe different possible values for the error term of a statistical model.

11.1 Assumptions About Error Terms

It's easy to write out an equation that includes an error term, but we are not going to be able to do much with our model unless we make some assumptions about the error term. One of the most important (and challenging) parts of doing statistical analysis is making assumptions

about the possible values of the error term. Different assumptions about the error term can result in very different conclusions.

Let's again consider the simple model of happiness that was introduced above:

$$happiness = 3.0 + 2.3 \times income + \varepsilon$$

We might assume the following things about the error term (ε):

1. The values of the error term (ε) can be described by a normal distribution with a mean of 0
2. Knowing someone's income doesn't help us predict the values of the error term (ε)

What do these two assumptions mean?

First, if the error term (ε) follows a normal distribution with a mean of zero, that means that (according to our model), people are just as likely to have a positive value of the error term as they are to have a negative value of the error term. In other words, all those factors we haven't accounted for in our model are equally likely to push people in the direction of being happier or in the direction of being less happy. Our model and assumptions tell us that if we predict happiness purely based on income, we'll *overestimate* some people's happiness, and we'll *underestimate* an equal number of people's happiness.[1]

Second, these assumptions allow us to describe how much individual observations will tend to deviate from our income-based predictions. We haven't specified in our assumptions what the standard deviation is for the normal distribution for the error term (ε), but statistical analysis will let us estimate the standard deviation of an error term. And we know that there is a 95% chance of drawing a value within two standard deviations of the mean for any normal distribution. So whatever the standard deviation of the error term (ε) is, we would expect that 95% of the time, the error term will take on a value that is within two standard deviations of zero. Conversely, 5% of the time, the error term will take on a value that is more than two standard deviations from zero. Suppose that the standard deviation of the error term (ε) happens to be three. If we have a dataset containing the income and happiness of 1,000 randomly selected people, we would expect that about 950 of these people will have a level of happiness that falls within six units of our income-based prediction. But for about 50 of these people, our prediction of their happiness will be off by more than six units.

Third, our assumptions imply that income is not tied in any consistent way to (the total sum of) factors other than income that also affect peoples' happiness. Remember, the error term (ε) represents all factors other than income that affect satisfaction. If income is related to these other factors, then the value of income should help us predict the value of the error term. For example, if having a stable environment in childhood tends to cause both higher incomes

[1] Note that these deviations from our prediction don't imply that our model is wrong; our model explicitly acknowledges that we'll get only imperfect estimates if we predict happiness based on income, since the unobserved error term (ε) also contributes to happiness.

and greater happiness in adulthood, the error term will partially reflect the effect of childhood stability on happiness, so high incomes (which are partially caused by childhood stability) will be probably be predictive of a more positive error term. This would constitute a violation of our assumptions since we specifically indicated that income wasn't predictive of the error term. As this example illustrates, our assumptions about error terms are often quite strict, making it rather difficult in practice to build good models that account for uncertainty.

11.2 Models and Probabilistic Thinking

Despite the difficulty inherent in building models that accommodate uncertainty, we have little alternative unless we wish to only build models of things we think we can predict with 100% accuracy. And fortunately, our models do not always have to be perfectly correct in order to generate useful predictions or explanations. As the statistician George Box famously said, "all models are wrong, but some are useful."

An important part of learning to do good statistical analysis is learning to think clearly about models so that you can pick out a model that is useful for whatever it is you want to accomplish. And the first step toward understanding many statistical models is learning to think about the world in probabilistic terms, as we've done here in this reading. Probabilistic thinking asks questions like:

- Based on what I do know and what I don't know, what can I predict?
- How does adding or removing different pieces of information change my prediction?
- How much uncertainty is there in my prediction?
- How often will my prediction differ greatly from what actually happens (even if my model is correct)?

12 Regression with Qualitative Independent Variables

Let's say I'm interested in studying how personality relates to gender. The most common personality measure in psychology is called the "Big Five" personality inventory. There is a standard set of 50 survey items that researchers can use to measure five aspects of personality. Figure 12.1 is an example of some of these questions and how they are formatted:

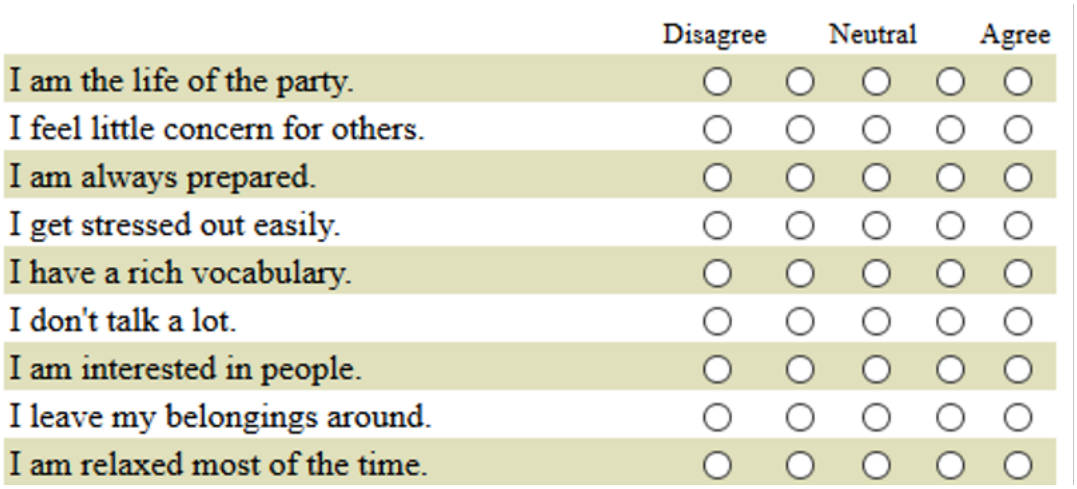
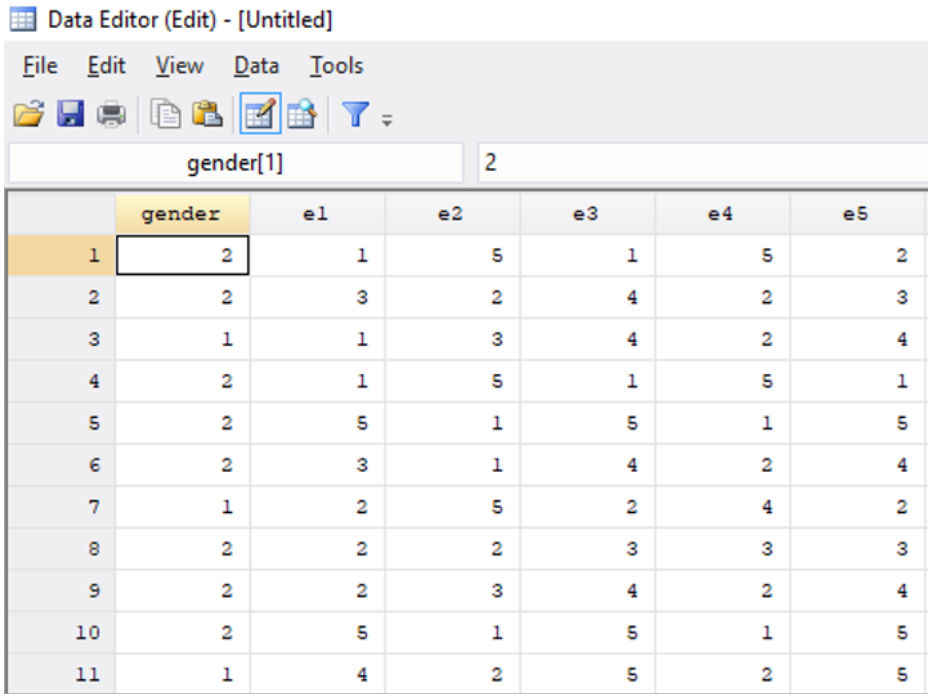


Figure 12.1

For now, I decide to focus on whether people are introverted or extroverted. Extroverts are outgoing and tend to enjoy interacting with others. Extroverts will tend to agree with the statement "I am the life of the party" while introverts will tend to agree with the item "I don't talk a lot."

I find a dataset that contains lots of responses to the Big Five personality questions as well as information on the gender of each respondent.[1] There are 10 different questions related to extroversion, and the dataset has one variable (column of data) for each of these 10 questions. The column labeled e1 shows responses to the item "I am the life of the party." A value of 1 means the respondent disagrees with this statement, while a 3 indicates neutral, and a 5 means they disagree.

[1] https://openpsychometrics.org/_rawdata/ (the file I used is called "BIG5.zip")



	gender	e1	e2	e3	e4	e5
1	2	1	5	1	5	2
2	2	3	2	4	2	3
3	1	1	3	4	2	4
4	2	1	5	1	5	1
5	2	5	1	5	1	5
6	2	3	1	4	2	4
7	1	2	5	2	4	2
8	2	2	2	3	3	3
9	2	2	3	4	2	4
10	2	5	1	5	1	5
11	1	4	2	5	2	5

Figure 12.2

For all of the odd-numbered extroversion questions (e1, e3, e5, etc.), agreement indicates extroversion. For the even-numbered items (e2, e4, e6, etc.), agreement indicates introversion. To create a single extroversion variable that combines responses from all 10 survey items, I create a tally, adding up all the values for odd-numbered questions and then subtracting the responses to the even-numbered questions. An extreme extrovert will have a 5 for all the odd-numbered questions and a 1 for all of the even-numbered ones, giving them a score of 20 (5x5-5x1=20). An extreme introvert will have a -20 since they will answer 1 to all the odd-numbered questions and 5 to all the even-numbered ones (5x1-5x5=-20).

Most people lie somewhere in the middle between introversion and extroversion:

Our gender variable was measured by asking respondents "What is your gender?" and they could choose from male, female, or other. In a moment, we'll consider those who responded "other," but for now, let's just look at those who chose either male or female.

12.1 Predicting extraversion using gender

If I want to describe differences in extraversion by gender in this dataset, I can compute the mean value of extraversion for males and for females. It turns out that males have an average extraversion of -0.46 while females' average level of extraversion is 0.53. Thus, the average

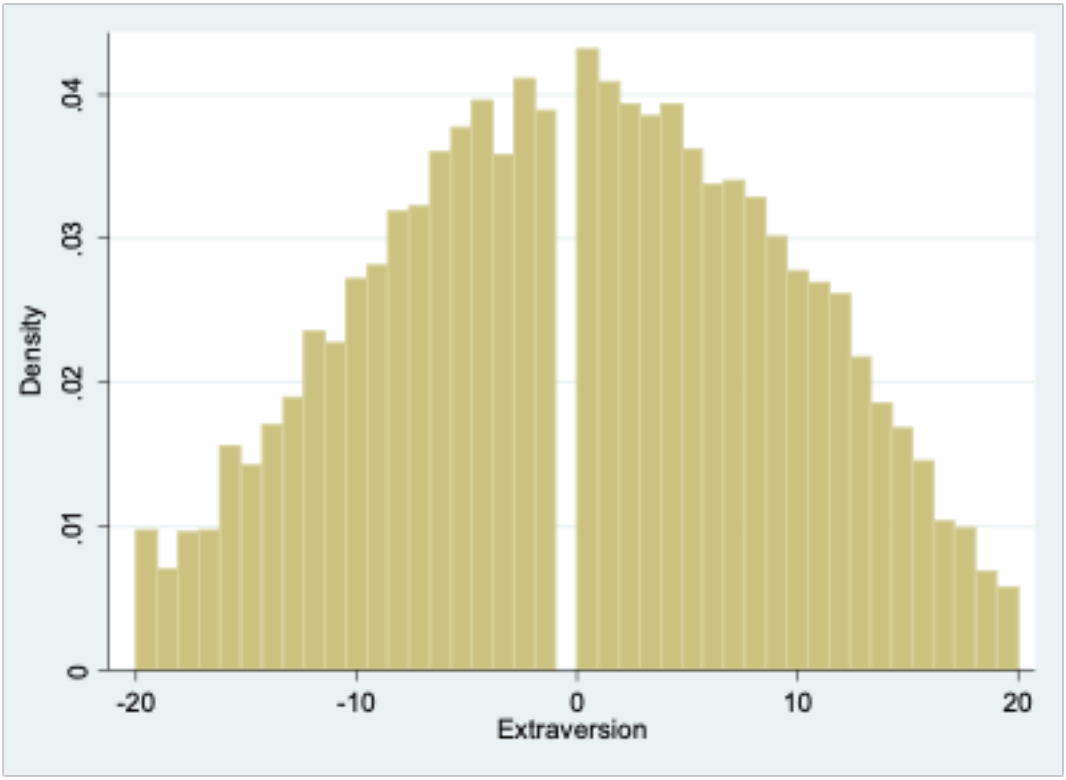


Figure 12.3

female is about 1-point more extraverted than the average male. But of course, there is lots of variation in extraversion among both groups:

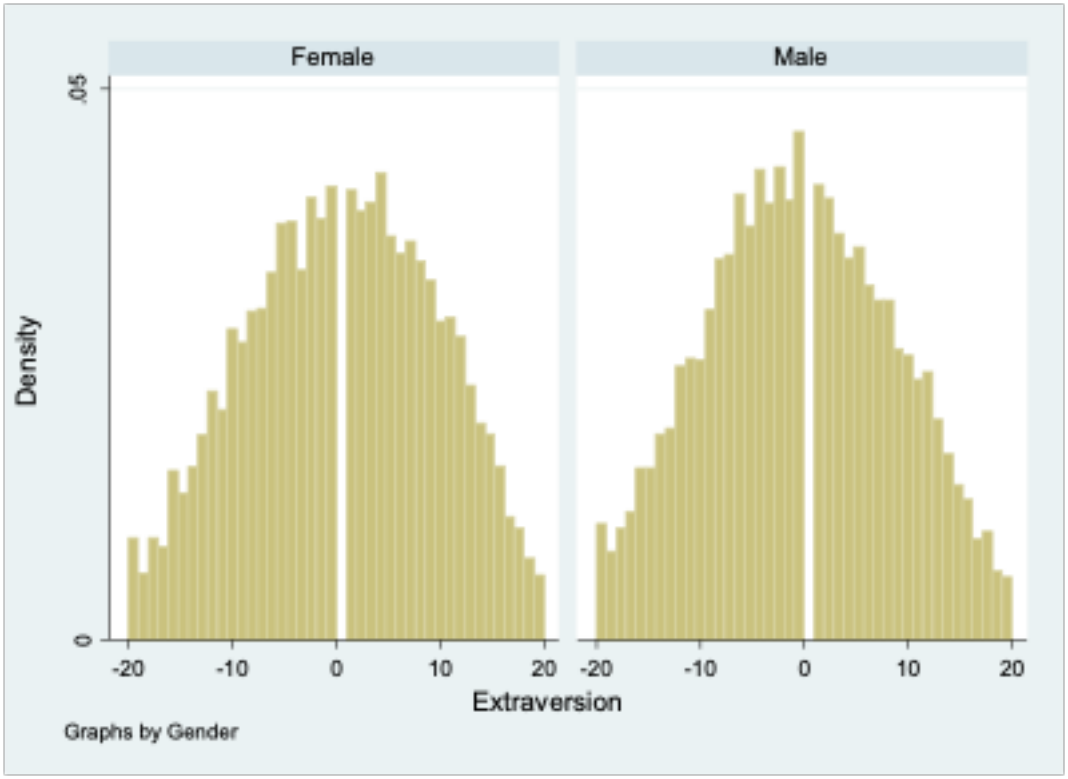


Figure 12.4

There are plenty of females who are introverts and plenty of males who are extroverts.

If you asked me to guess the extroversion level of someone and the only thing you told me about them was their gender, my best bet would probably be to guess the average extroversion level for someone of that gender. So for a female I knew nothing else about, I would guess their extroversion to be 0.53, while for a male I'd guess -0.46.

Social scientists use the **dependent variable** to describe the variable they're making a prediction about and **independent variable** to describe the variables that help them make that prediction. So in this example, extraversion is my dependent variable and gender is my independent variable.

When we're working with data, sometimes it's helpful to express how I would make a guess about a dependent variable (extraversion) based on other factors (gender) using a mathematical formula. In fact, this is exactly what we do when we run a regression. There are many ways I could write this formula, but I'll show just two for now. First, I could write:

$$\widehat{Extraversion} = 0.53 \times Female - 0.46 \times Male \tag{12.1}$$

Notice I've added a "hat" above the name of the variable *Extraversion*; this hat means that I'm making a guess about the value of that variable (I'm guessing the level of extraversion based on gender). The equation has two other variables *Female* and *Male*, and these two variables will take on a value of 1 if the person's gender is equal to the name of the variable and will otherwise take on a value of 0. For a female, *Female* will equal 1 and *Male* will equal 0, giving us:

$$\widehat{Extraversion} = 0.53 \times (1) - 0.46 \times (0) = 0.53$$

So our guess for the level of extroversion ($\widehat{Extraversion}$) of a female we know nothing about is 0.53.

For a male, our guess is:

$$\widehat{Extraversion} = 0.53 \times (0) - 0.46 \times (1) = -0.46$$

There's a second way I can write my formula, which will turn out to be more useful in the future when we come to consider multiple factors at the same time that might help us predict the value of a dependent variable. Rather than having two variables to represent gender in my equation, I can just use one:

$$\widehat{Extraversion} = 0.53 - 0.99 \times Male \tag{12.2}$$

In Equation 12.2, we start from female as our baseline. Notice that the first number we see (0.53) is our guess for the value of extraversion for a female. When we're considering a female, Male=0, so:

$$\widehat{Extraversion} = 0.53 - 0.99 \times 0 = 0.53$$

Thus, we get the right prediction for females from this equation, even though we didn't include a variable specifically for females. If we have a male, Male=1, so we get:

$$\widehat{Extraversion} = 0.53 - 0.99 \times 1 = -0.46$$

This is the same prediction we got before. Remember, I decided to initially just analyze respondents who selected either male or female. Since we are only considering two categories (male or female), and each respondent is either a male or a female, saying $Male = 1$ lets me know that $Female = 0$. It's actually repetitive in this context to both say that $Male = 1$

and *Female* = 0. Similarly, saying *Male* = 0 implies that *Female* = 1. So I can simplify my equation by just including one variable to indicate binary gender.

Notice that in Equation 12.2, the number next to *Male* is equal to the difference between the average level of extraversion for females and the average level for males (0.53-(-0.46)=0.99). This is because Equation 12.2 starts with females as the baseline, so to get our prediction for males, we have to adjust our baseline prediction by the average difference for males.

Equation 12.2 is also typically how we will arrange our equation when we're running a regression.

12.2 Prediction with more than two categories for gender

I now move beyond the gender binary and consider the "other" category in survey responses. I'll refer to this other category as "non-binary" gender. The average level of extraversion among those with non-binary gender is -5.66. So non-binary people tend to be quite a bit more introverted than those who identify as male or female. As with males and females, there is considerable variation among non-binary people:

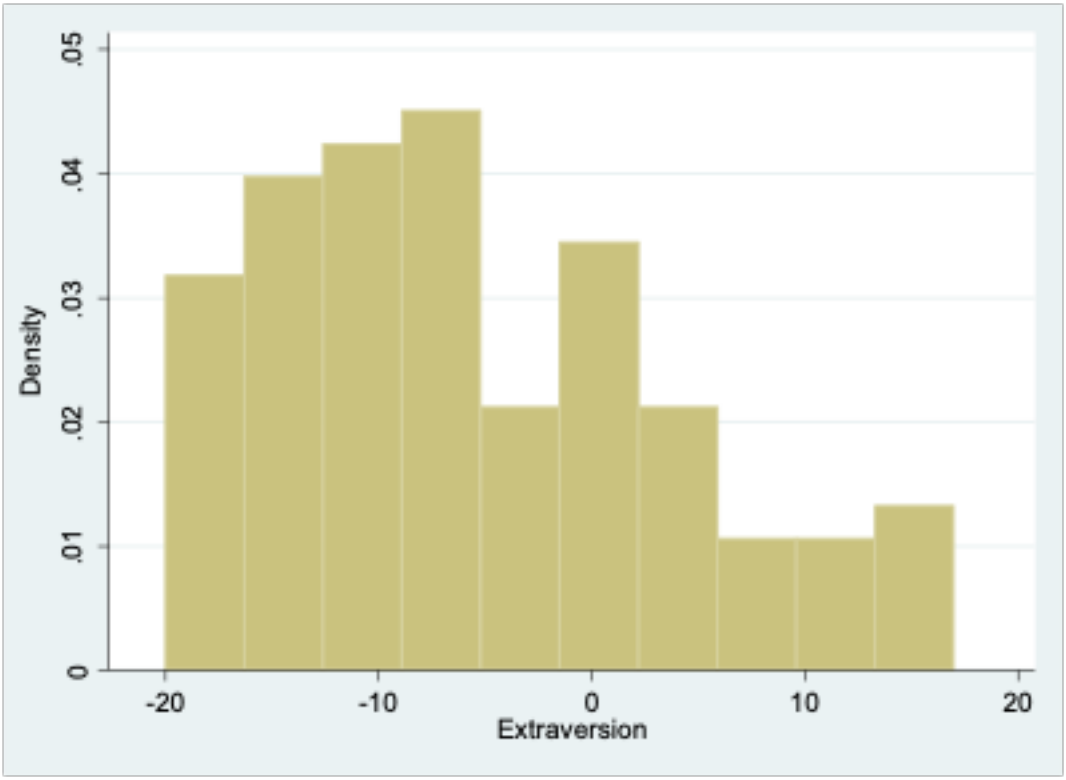


Figure 12.5

The number of non-binary respondents is relatively small (102), so it's not terrible surprising that this histogram looks a bit choppier than the ones we saw before.

Again, if we had to make a guess about the level of extraversion of someone, and all we knew about that person was that their gender was non-binary, we would probably want to guess the mean value among non-binary respondents (-5.66). Modifying Equation 12.1 to incorporate a third category is relatively straightforward:

$$\widehat{Extraversion} = 0.53 \times Female - 0.46 \times Male - 5.69 \times Other \tag{12.3}$$

For someone who identifies as female, we would plug in $Female = 1$, $Male = 0$, and $Other = 0$:

$$\widehat{Extraversion} = 0.53 \times (1) - 0.46 \times (0) - 5.66 \times (0) = 0.53$$

If someone identifies as non-binary, we would use $Female = 0$, $Male = 0$, and $Other = 1$:

$$\widehat{Extraversion} = 0.53 \times (0) - 0.46 \times (0) - 5.66 \times (1) = -5.66$$

We can also return to the format of Equation 12.2 but modify it to include the other category. This is how we will typically write our equation if we are doing a regression:

$$\widehat{Extraversion} = 0.53 - 0.99 \times Male - 6.19 \times Other \tag{12.4}$$

Now that there are three possible values for gender (female, male, and other), knowing the value of *Male* doesn't necessarily allow us to conclude what the vale of female is. If , the individual could identify as either female or non-binary. So we have to include a second variable. In this case, we chose to include the variable *Other*. If we know the values of *Male* and *Other*, we can always figure out the value of *Female* by process of elimination.

For a non-binary person, we plug in $Male = 0$, and $Other = 1$:

$$\widehat{Extraversion} = 0.53 - 0.99 \times (0) - 6.19 \times (1) = -5.66$$

When considering a female, we use $Male = 0$, and $Other = 0$:

$$\widehat{Extraversion} = 0.53 - 0.99 \times (0) - 6.19 \times (0) = 0.53$$

Equation 12.3 and Equation 12.4 communicate an equivalent method of making a prediction about extraversion based on gender; they just offer this information in two different formats.

Equation 12.4 might be a bit trickier to understand for now, but it will become very useful in the future.

Notice that we can talk about gender either as one qualitative variable with three possible values (female, male, or other), or we can talk about it as a series of three dummy variables (*Female*, *Male*, and *Other*) that can take each on a value of either 0 or 1. This can make things a bit confusing, but the important thing to remember is that when we have a qualitative variable with more than two categories, we'll need to break out the categories into a set of dummy variables for purposes of representing the qualitative variable in an equation.

However, as Equation 12.2 and Equation 12.4 illustrate, we don't necessarily need a dummy variable for every single category. Specifically, whenever we want to create an equation with a qualitative independent variable in a format like Equation 12.2 or Equation 12.4, the number of dummy variables should be equal to the number of categories minus one. Since our gender variable can take on three possible values in this example, we included two independent variables in Equation 12.4. No dummy variable is included for female, so we call female the **omitted category** or the **baseline category**. Remember, the first number in Equation 12.4 is 0.53, which represents our guess for females—the baseline category. If we instead had a qualitative variable with five categories, we would include four dummy variables in our equation.

13 Regression with Qualitative Dependent Variables

Suppose I want to build a model of voting. I decide to use the 2016 American National Election Studies[1] survey results to try to understand how race is associated with voting. Respondents in the 2016 survey were asked about who they voted for in 2012, and I'm going to focus on their 2012 voting patterns for now. Here are the distributions for my two main variables of interest:

```
```{stata}

. tab vote
PRE: RECALL OF LAST (2012) PRESIDENTAL  |
                            VOTE CHOICE |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
                        1. Barack Obama |      1,728       56.58       56.58
                         2. Mitt Romney |      1,268       41.52       98.10
                       5. Other SPECIFY |         58        1.90      100.00
----------------------------------------+-----------------------------------
                                  Total |      3,054      100.00

. tab race
  PRE: SUMMARY - R SELF-IDENTIFIED RACE |      Freq.     Percent        Cum.
----------------------------------------+-----------------------------------
                1. White, non-Hispanic |      3,038       71.68       71.68
                2. Black, non-Hispanic |        398        9.39       81.08
3. Asian, native Hawaiian or other Paci |        148        3.49       84.57
4. Native American or Alaska Native, no |         27        0.64       85.21
                           5. Hispanic |        450       10.62       95.82
6. Other non-Hispanic incl multiple rac |        177        4.18      100.00
----------------------------------------+-----------------------------------
                                  Total |      4,238      100.00
```
```

[1] https://electionstudies.org/data-center/2016-time-series-study/

Notice that my dependent variable (vote) is qualitative. It can take on three possible values: voted for Obama, voted for Romney, or voted for other. I can build a simple set of regression models to see how race predicts vote choice. The key is to first convert each of the three categories for my dependent variable into its own dummy variable. I can accomplish this with the following code:

```{stata}
tab vote, gen(vote_)
```

I now have several new variables in my dataset that have names starting with "race_":

```{stata}
. tab vote_1

   vote==1. |
     Barack |
      Obama |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,326       43.42       43.42
          1 |      1,728       56.58      100.00
------------+-----------------------------------
      Total |      3,054      100.00

. tab vote_2

   vote==2. |
Mitt Romney |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,786       58.48       58.48
          1 |      1,268       41.52      100.00
------------+-----------------------------------
      Total |      3,054      100.00

. tab vote_3

   vote==5. |
      Other |
    SPECIFY |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,996       98.10       98.10
          1 |         58        1.90      100.00
```

```
------------+-----------------------------------
      Total |      3,054      100.00
```

I also convert my race variable into a set of dummy variables by running:

```{stata}
tab race, gen(race_)
```

I can then run three regressions, one for each value of my dependent variable. Let's start with voting for Obama (vote_1):

```{stata}
. reg vote_1 race_2 race_3 race_4 race_5 race_6

      Source |       SS           df       MS      Number of obs   =     3,036
-------------+----------------------------------   F(5, 3030)      =     76.29
       Model |  83.3981974         5  16.6796395   Prob > F        =    0.0000
    Residual |  662.426572     3,030  .218622631   R-squared       =    0.1118
-------------+----------------------------------   Adj R-squared   =    0.1104
       Total |  745.824769     3,035  .245741275   Root MSE        =    .46757

------------------------------------------------------------------------------
      vote_1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      race_2 |   .4972868   .0281049    17.69   0.000     .4421802    .5523934
      race_3 |   .2078207   .0541766     3.84   0.000     .1015941    .3140472
      race_4 |   .1028423   .1353307     0.76   0.447     -.162507    .3681916
      race_5 |   .3135004    .032158     9.75   0.000     .2504466    .3765542
      race_6 |   .1042547   .0441427     2.36   0.018      .017702    .1908075
       _cons |    .480491   .0097901    49.08   0.000     .4612952    .4996868
```

Since our independent variable is qualitative, we have an omitted category. In this case, we've left category 1 (race_1) out of our regression, which indicates non-Hispanic white respondents. Our constant indicates that predicted value of the dependent variable when all independent variables are equal to zero. We can see this by writing out the regression equation:

$$\widehat{vote\_1} = .48 + .50race\_2 + .21race\_3 + .10race\_4 + .31race\_5 + .10race\_6 \qquad (13.1)$$

For non-Hispanic white respondents, race_1 equals one and all other race dummy variables equal zero, so we get:

$$\widehat{vote\_1} = .48 + .50(0) + .21(0) + .10(0) + .31(0) + .10(0) = .48$$

Remember, vote_1 is equal to zero if the respondent didn't vote for Obama, and it is equal to one if the respondent did vote for Obama. Our predicted value is neither zero nor one; instead, we get .48. This can be interpreted as indicating the probability of a one. In other words, a non-Hispanic white has a .48 probability of voting for Obama. We can also convert this probability to a percentage by moving the decimal place two spots to the right: a non-Hispanic white is estimated to have a 48% chance of voting for Obama, according to this model.

Now, let's look at the slope coefficients. The coefficient for black (race_2) equals .50. Thus, a one-unit increase in race_2 is associated with a .50-unit increase in vote_1. Let's break that down a bit to see if we can create a clearer interpretation. Since race_2 is a dummy variable and non-Hispanic white is the omitted category, a one-unit increase in race_2 correspondents to having a black respondent instead of a white respondent. And since our dependent variable is binary, we should think in terms of probabilities, which can be converted to percentages: a .50-unit increase in vote_1 means a 50 percentage-point increase in the probability of voting for Obama. So putting this altogether, we'd say: (non-Hispanic) black voters are 50 percentage points more likely to vote for Obama than (non-Hispanic) white voters, according to this model.

Similarly, Asian voters are 21 percentage points more likely to vote for Obama than (non-Hispanic) white voters. Native Americans are 10 percentage points more likely to vote for Obama than (non-Hispanic) white voters. Hispanics are 31 percentage points more likely to vote for Obama than non-Hispanic white voters. And voters identifying as multiracial or other race are 10 percentage points more likely to vote for Obama than (non-Hispanic) white voters. All of these differences are statistically significant, except for Native American versus white voters (probably because there are only 27 Native Americans in the sample, making the estimate of this difference very imprecise).

Let's move onto running a regression for the second category of our dependent variable:

```
```{stata}
. reg vote_2 race_2 race_3 race_4 race_5 race_6

      Source |       SS           df       MS      Number of obs   =     3,036
-------------+----------------------------------   F(5, 3030)      =     72.35
       Model |  78.6117037         5  15.7223407   Prob > F        =    0.0000
    Residual |  658.463395     3,030  .217314652   R-squared       =    0.1067
-------------+----------------------------------   Adj R-squared   =    0.1052
```
```

```
       Total |  737.075099     3,035  .242858352   Root MSE        =    .46617

------------------------------------------------------------------------------
      vote_2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      race_2 |   -.483031   .0280207   -17.24   0.000    -.5379725   -.4280895
      race_3 |  -.2002027   .0540143    -3.71   0.000     -.306111   -.0942944
      race_4 |  -.0822373   .1349253    -0.61   0.542    -.3467917     .182317
      race_5 |  -.3014791   .0320617    -9.40   0.000     -.364344   -.2386142
      race_6 |  -.1344972   .0440105    -3.06   0.002    -.2207906   -.0482038
       _cons |    .498904   .0097607    51.11   0.000     .4797657    .5180423
```

Now we're looking at predictions of voting for Mitt Romney. Our constant is .50, indicating that a non-Hispanic white voter has a 50% chance of voting for Mitt Romney. The coefficient of -.48 for race_2 indicates that (non-Hispanic) black voters are 48 percentage points less likely to vote for Mitt Romney than (non-Hispanic) white voters. I won't go on to interpret the rest of the coefficients, but they follow the same pattern.

Finally, let's look at a regression with vote_3 as the dependent variable:

```
```{stata}
. reg vote_3 race_2 race_3 race_4 race_5 race_6

      Source |       SS           df       MS      Number of obs   =     3,036
-------------+----------------------------------   F(5, 3030)      =      2.23
       Model |   .20833556         5  .041667112   Prob > F        =    0.0490
    Residual |  56.6836275     3,030  .018707468   R-squared       =    0.0037
-------------+----------------------------------   Adj R-squared   =    0.0020
       Total |  56.8919631     3,035  .018745293   Root MSE        =    .13678

------------------------------------------------------------------------------
      vote_3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      race_2 |  -.0142558   .0082213    -1.73   0.083    -.0303757    .0018642
      race_3 |   -.007618   .0158479    -0.48   0.631    -.0386917    .0234557
      race_4 |   -.020605   .0395873    -0.52   0.603    -.0982258    .0570158
      race_5 |  -.0120213    .009407    -1.28   0.201     -.030466    .0064234
      race_6 |   .0302425   .0129128     2.34   0.019     .0049238    .0555611
       _cons |    .020605   .0028638     7.19   0.000     .0149898    .0262202
```
```

This regression provides some insights into who supported third-party candidates in the 2012 election. First, our constant indicates that a non-Hispanic white voter has a 2% chance of voting third-party. (Non-Hispanic) black voters are one percentage point less likely to vote third-party than white voters, although this difference is only significant at the .10 level. The only other significant slope coefficient is for race_6, where we see that people who identify as multiracial or other race are estimated to be three percentage points more likely to vote third-party than (non-Hispanic) white respondents.

Now that we've run one regression for each category of our dependent variable, we've completed an analysis. Note that using regular linear regression (the reg function in Stata) is not the only way (or even necessarily the preferred way) to analyze a qualitative dependent variable. There are other models (e.g., multinomial logistic regression) that are specifically designed to be used with a qualitative dependent variable. However, using simple linear regression is a good way to get started looking at qualitative variables if you haven't learned these fancier models and how to properly interpret them.

One final thing I want to show you is that our results will be in a slightly different format but will be in one sense equivalent if we decide to use a different category as our omitted category when using a qualitative independent variable. Let's say we want to make black (race_2) our reference category. Compare the following results to what we saw near the top of this page:

```{stata}
. reg vote_3 race_1 race_3 race_4 race_5 race_6

      Source |       SS           df       MS      Number of obs   =     3,036
-------------+----------------------------------   F(5, 3030)      =      2.23
       Model |   .20833556         5  .041667112   Prob > F        =    0.0490
    Residual |  56.6836275     3,030  .018707468   R-squared       =    0.0037
-------------+----------------------------------   Adj R-squared   =    0.0020
       Total |  56.8919631     3,035  .018745293   Root MSE        =    .13678

------------------------------------------------------------------------------
      vote_3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      race_1 |   .0142558   .0082213     1.73   0.083    -.0018642    .0303757
      race_3 |   .0066378    .017388     0.38   0.703    -.0274557    .0407313
      race_4 |  -.0063492   .0402287    -0.16   0.875    -.0852274     .072529
      race_5 |   .0022345   .0118186     0.19   0.850    -.0209387    .0254077
      race_6 |   .0444983   .0147623     3.01   0.003      .015553    .0734435
       _cons |   .0063492   .0077064     0.82   0.410    -.0087611    .0214595
```

Now, our constant tells us that a black voter has a .6% chance of voting third-party. This is the same prediction we would get from our prior model where race_1 was the omitted category: to

find our prediction for black voters from the prior results we would have added the coefficient for race_2 (-.014) to the constant (.021), yielding .6% or .006 (or .007 if we use the rounded numbers shown in parentheses).

The coefficient for race_1 tells us about how white voters differ from black voters. Notice that the p-value is exactly the same as what we saw in the prior table for race_2, and the coefficient for race_1 in this table is the same as the coefficient for race_2 in the prior table, except the sign has changed. That's because comparing black to white is the same as comparing white to black, except that we're going in the opposite direction.

You can go on to play around with these two sets of results more on your own if you'd like. Both regression equations will yield the same prediction for a voter of any given race. The difference lies only in the starting point, as represented fby the constant. However, the p-values will usually differ because they are describing a different comparison (e.g., comparing Asian to black in this table versus comparing Asian to white in the prior table). Thus, it doesn't really matter which category you pick as your omitted category, except that you may care more about some comparisons than others. You can also run the same regression multiple times but with different omitted categories so that you can get the p-values for a full set of comparisons across groups.