Chapter 6 – Analyzing Bivariate Data

Introduction

https://bit.ly/probstatsUnit6
(4 sections in this unit)

In chapter 5 we learned how to analyze and describe univariate, or single-variable data. We explored ways to present our data visually with graphs and charts and how to analyze our data with numerical statistics. Also, we described our findings verbally and in context. Now we will be analyzing bivariate numerical data. This means that two numerical values have been collected about each individual. Such bivariate data is often given in a table or listed as ordered pairs. We will construct appropriate graphs, calculate numerical statistics and equations, and describe the relationship between the two variables in context. The purpose will be to explore whether or not a relationship or association exists between the two numerical variables. If an association does exist, statistics can be used to predict one variable based on the other variable.

6.1 Displaying Bivariate Data

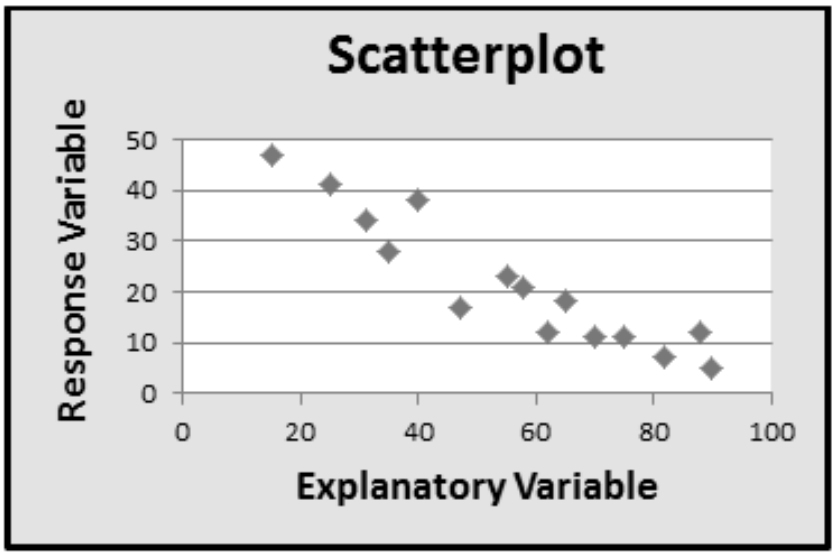
Learning Objectives

- Construct and interpret scatterplots
- Identify explanatory and response variables
- Describe bivariate distributions in context—including strength, outliers, form and direction

Scatterplots

Scatterplots are graphs that represent a relationship between two variables. Two numerical values are measured about each individual being studied. When these two values become ordered pairs that are graphed on a coordinate plane, the resulting graph is called a **scatterplot**. We often suspect that one of these variables might explain, cause changes in, or help to predict the other variable. The **explanatory variable** is the variable that we believe may explain or affect the other variable. The explanatory variable is plotted along the x-axis. The **response variable** is the variable we believe may respond to, or be affected by, the explanatory variable. The response variable is plotted along the y-axis. The explanatory variable is often referred to as the independent variable and the response variable is referred to as the dependent variable. Even though we often look for an explanatory-response relationship between the two variables, we can create a scatterplot even if no such relationship exists.

https://bit.ly/probstatsSection6-1
(4 videos in this section)



Example 1

State whether or not you suspect that there will be an explanatory-response relationship between each of the following pairs of data. If yes, identify the explanatory and response variables.

a) A college professor decided to examine whether or not there is a relationship between the amount of time that a student studies and his or her score on the mid-term exam. At the end of the exam, each student was asked to record the number of hours he or she had spent studying for the mid-term. The professor then made a scatterplot to examine the data.

b) A different professor wanted to see whether or not there is an association between her students' heights and their IQ scores. She gave each of her students an IQ test and had her TA (teaching assistant) measure each student's height to the nearest inch. She constructed a scatterplot to examine the data.

Solution

a) It is reasonable to believe that the amount of studying does somehow have an effect on students' exam scores. The explanatory variable is hours studying and the response variable is exam score. Often thinking in terms of a cause and effect relationship can help identify which variable is which. As a hint, try to determine if one of the variables comes first. If one variable does happen first, then it is most likely the explanatory variable. In our example, studying should come before the exam.

b) It is not reasonable to believe that there is an association between height and IQ scores. Neither of these variables comes before the other and neither would be useful in predicting the other. However, even though we do not believe that there is an explanatory-response relationship between these variables, we can still construct a scatterplot

Example 2

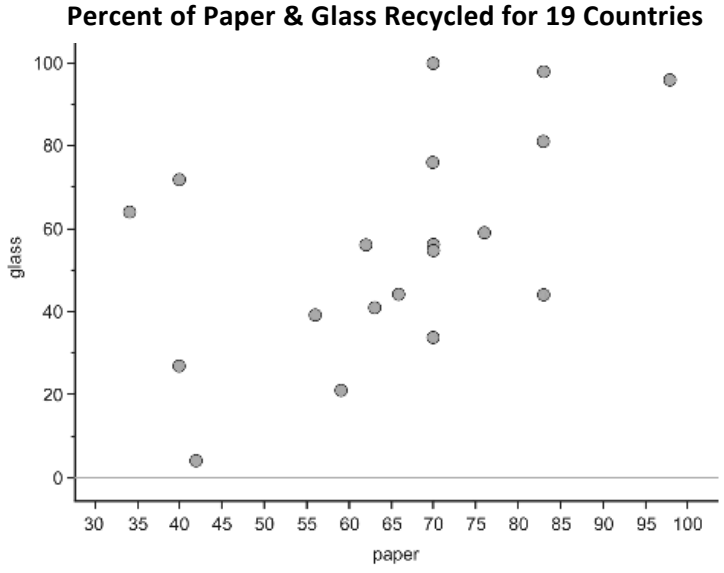
The following table reports the recycling rates for paper and glass packaging for several individual countries. It would be interesting to see if there is a predictable relationship between the percentages of each material that countries recycle. Construct a scatter plot to examine the relationship. Treat percentage of paper packaging recycled as the explanatory variable.

Country	% of Paper Packaging Recycled	% of Glass Packaging Recycled
Estonia	34	64
New Zealand	40	72
Poland	40	27
Cyprus	42	4
Portugal	56	39
United States	59	21
Italy	62	56
Spain	63	41
Australia	66	44
Greece	70	34
Finland	70	56
Ireland	70	55
Netherlands	70	76
Sweden	70	100
France	76	59
Germany	83	81
Austria	83	44
Belgium	83	98
Japan	98	96

Figure: Paper and Glass Packaging Recycling Rates for 19 countries

Solution

We will place the paper recycling rates on the horizontal axis because we are treating it as the explanatory variable. Glass recycling rates are then plotted along the vertical axis. Next, plot a point that shows each country's rate of recycling for the two materials. Be sure to label your axes.



Notice that we do not always need to start at zero on either axis when making scatterplots.

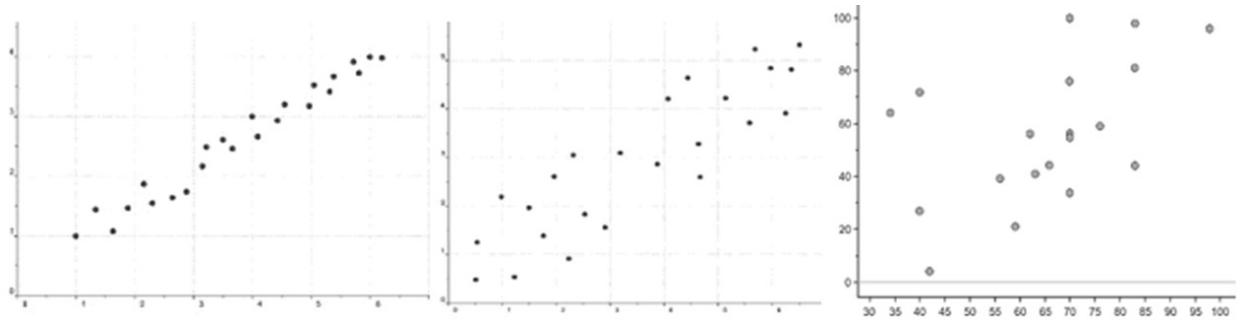
Describing Bivariate Data

When we describe single variable data, we address several characteristics. We used the acronym ***S.O.C.C.S.*** to help us to remember to describe the **s**hape, **o**utliers, **c**enter, **c**ontext and **s**pread of a distribution. For bivariate situations, we will again need to remember to discuss several key characteristics of the data. The important characteristics to describe when looking at the relationship between two numerical variables will be strength, outliers, form and direction. We will do this in the context of the variables and individuals being compared. The acronym that will help us to remember what to include in our descriptions is: **S.C.O.F.D.** (**S**trength, **C**ontext, **O**utliers, **F**orm and **D**irection).

When looking at a scatterplot, it is helpful to imagine drawing a **line-of-best-fit** through the data. A line-of-best-fit is a line that follows the trend of the data. It may go through some, all, or none of the actual points on the scatterplot. Do not actually draw such a line on your plot - just try to determine whether or not such a line would make sense, and if so, where it would fit. As you observe a scatterplot and imagine drawing such a line, you can ask yourself questions such as the following: *How close to a line do the points lie? Would a curved pattern fit better? Are there points that would be far away from the line? Would the line have a positive or negative slope?*

Strength

Once you have constructed a scatterplot, you can examine the strength of the relationship between the two variables. The **strength** refers to how closely the points form a pattern. The more closely the points fit a pattern, the stronger the relationship is between the variables. The more spread out and scattered the points are, the weaker the relationship. The first plot shows an extremely strong, linear pattern because the points form an obvious line. The second plot is more scattered so it is only moderately strong. The third plot does not show much of a pattern at all so it would be considered to have a weak association. Keep in mind that the association may be very strong, but not linear. We could find a very clear curved pattern in the data, for example. In the next section of this chapter we will learn about a statistic, called correlation, that measures the strength of the linear relationship between two variables.



VERY STRONG **MODERATELY STRONG** **VERY WEAK**

In Example #2, the strength of the relationship between paper and glass recycling rates for these countries would be considered weak.

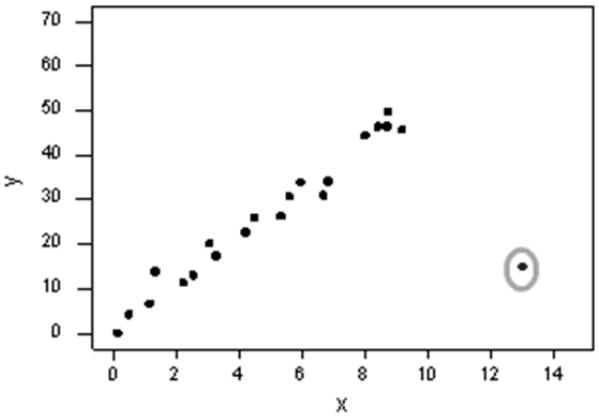
Context

Do not forget that the graph, the numerical values, and any equations, are all about some particular situation. Keep the **context** in mind when considering any bivariate situation. All of these elements should be described in the context of the variables and the individuals being examined. These graphs and statistics are not meaningless, they are about something!

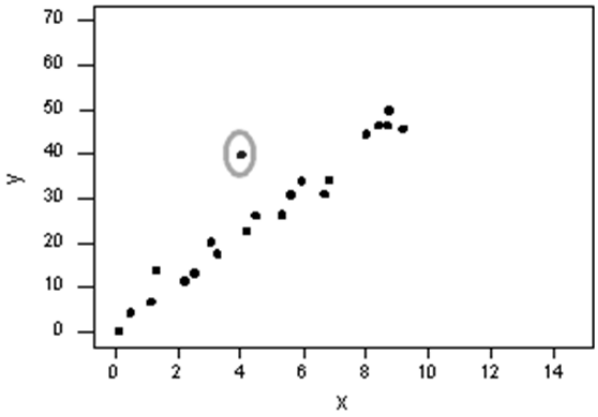
> *In Example #2, the scatterplot explores the relationship between glass and paper recycling rates for several countries.*

Outliers

When examining a scatterplot, look for any data values that do not fit the overall pattern of the rest of the data. An **outlier** will be a point that lies away from the rest of the data or one that seems to affect the strength of the relationship between the two variables. Some outliers will weaken the association between the variables. However, they often will not significantly change where a line-of-best-fit would be drawn. An **influential point** is an outlier that actually seems to influence the location of the line-of-best-fit. Imagine what the plot would look like without the point in question. If it would change the strength, then the point is an outlier. If it would change the slope of a line-of-best-fit, or where the line would be drawn, then the point is influential.



OUTLIER & INFLUENTIAL

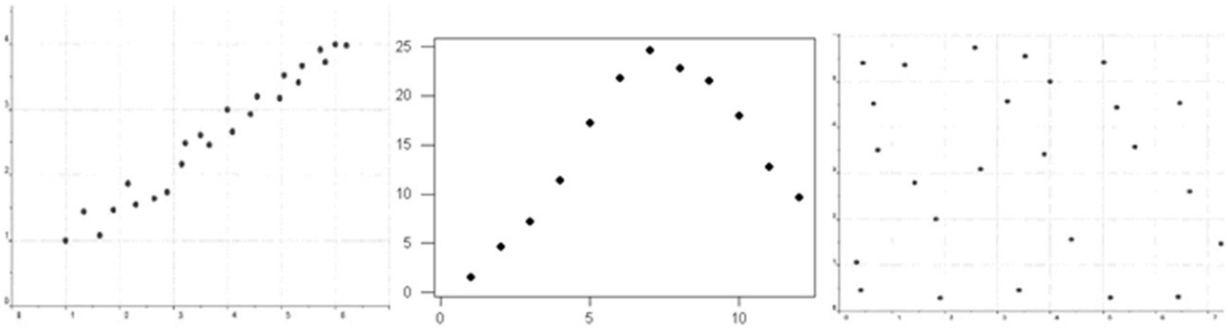


OUTLIER (but not influential)

> *In Example #2, there seem to be some outliers. For example, Estonia and New Zealand each have a much lower paper recycling rate than their glass recycling rate. Without these data values, the relationship in the rest of the scatterplot would be stronger.*

Form

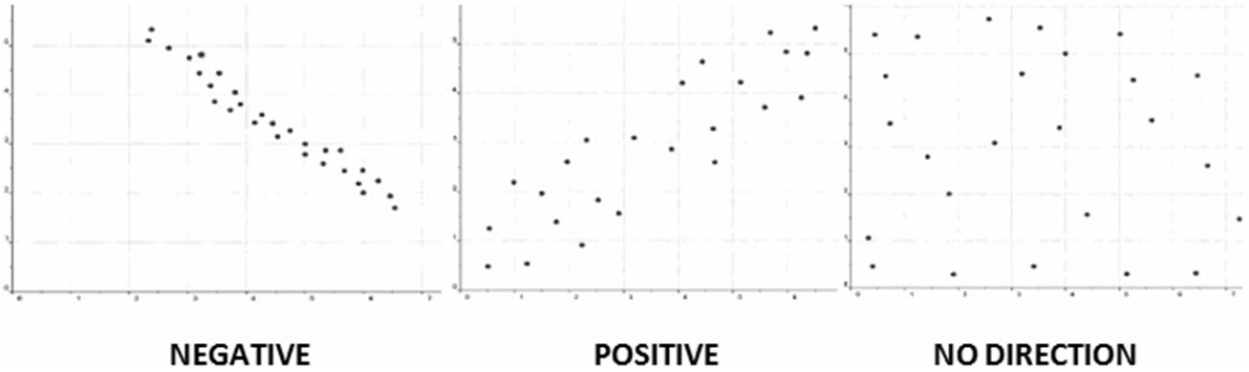
Many scatterplots show a clear form or pattern. The first plot below shows a clearly linear pattern or form. It is easy to imagine drawing a line-of-best-fit through these points. The second plot shows a clearly curved form. A line would not make any sense, so this is non-linear. The third plot shows a great deal of scatter among the points with no clear pattern. It has no form whatsoever.



In example #2, the scatterplot for paper and glass recycling rates shows a very weak linear form. The relationship is very weak but we can still see in general, that as paper recycling rates increase, glass recycling rates also increase. If the two outliers were removed, the scatterplot would have a stronger linear pattern.

Direction

The direction of the graph is also important to mention. A graph that goes down to the right has a **negative association**. As the explanatory variable increases, the response variable decreases. The first plot below or on next page has a negative relationship between the variables. A graph that goes up to the right has a **positive association**. As the explanatory variable increases, the response variable also increases. The second plot shows a positive relationship between the variables. The third plot is an example of a graph that has neither a positive, nor a negative direction. If the relationship is linear and a line-of-best-fit is added to the graph, the slope of the line will be positive if the association is positive. Likewise, the line will have a negative slope if there is a negative linear association between the two variables.

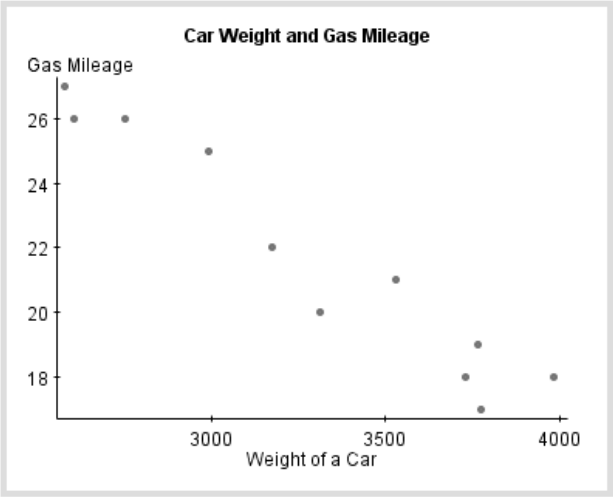


In example #2, the scatterplot for paper and glass recycling rates shows a positive association. As the paper recycling rate for these countries increases, so does the glass recycling rate.

S.C.O.F.D

When you describe the relationship between bivariate data there are several characteristics to include. The acronym **S.C.O.F.D.** will help you remember to describe the strength of the relationship, keep your description is in context, mention any outliers, describe the form, and state the direction of the graph.

Example 3



Consider the scatterplot to the right that shows the weights (pounds) and gas mileages (miles per gallon) for several cars.

a) Identify the explanatory and response variables.

b) Describe what the scatterplot shows. Be sure to address strength, context, outliers, form, and direction (S.C.O.F.D.).

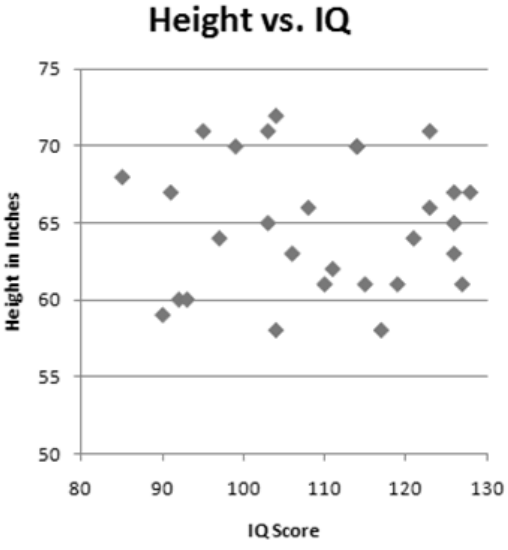
Solution

a) **The explanatory variable is** the weight of the cars in pounds.
The response variable is the gas mileage of the cars in miles per gallon.

b) The relationship between these vehicles' weights in pounds and gas mileage (mpg) is strong, negative, and linear. There are no clear outliers visible in the graph. As the weights of the vehicles increase, the gas mileages of the vehicles decrease.

Example 4

The scatterplot to the right shows the data collected by the professor who wanted to see whether or not there is an association between her students' heights and their IQ scores. She gave each of her students an IQ test and had her TA measure each student's height to the nearest inch. Describe what the scatterplot shows. Be sure to address strength, context, outliers, form, and direction (S.C.O.F.D.).



Solution

There appears to be no relationship between height and IQ scores for these students. The graph has no form and no clear direction. Therefore, there are no outliers. The relationship has no strength. There is no pattern or trend between IQ scores and students' heights.

Problem Set 6.1

Exercises

1) State whether or not you suspect that there will be an explanatory-response relationship between each of the following pairs of data. If yes, identify the explanatory and response variables.

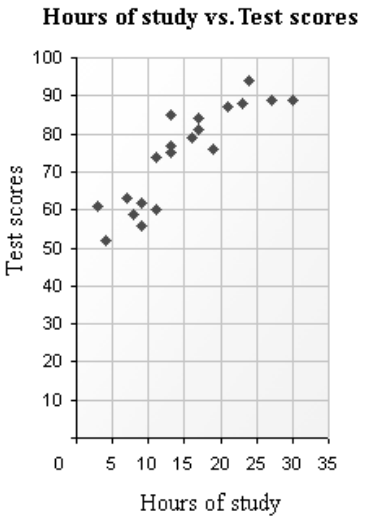
   a) The number of semesters that students have been enrolled in college and the number of credits that they have earned.

   b) Student grades on a statistics test and their weights.

   c) Employees' annual salaries and the number of years that the employee has been employed by the company.

   d) The number of applications downloaded on a student's cell phone and the number of months that they have owned the cell phone.

2) A college professor decided to examine whether or not there is a relationship between the amount of time that a student studies and his or her score on the mid-term exam (out of 100 points possible). At the end of the exam each student was asked to record the number of hours he or she had spent studying for the mid-term. The professor then made the scatterplot shown to the right to examine the data. Describe what the scatterplot shows. Be sure to address strength, context, outliers, form, and direction (S.C.O.F.D.).



3) Malia turned the water on in her bathtub full blast. She then measured the depth of the water every two minutes until the bathtub was full (and her mother started to freak out). Her findings are listed in the table to the right.

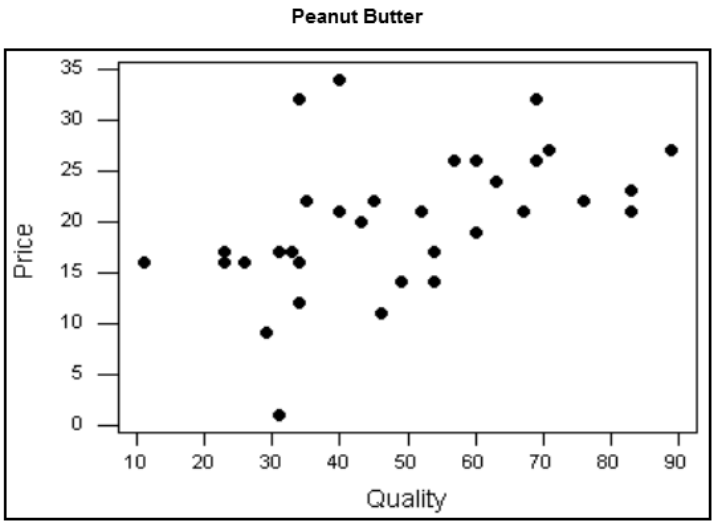
   a) Identify the explanatory and response variables for this situation.

   b) Construct a scatterplot to show the results.

   c) Describe what the scatterplot shows. Be sure to address strength, context, outliers, form and direction (S.C.O.F.D.).

Time (minutes)	Depth (cm)
2	7
4	9.5
6	14
8	19.5
10	21
12	24
14	32
16	36
18	37.5
20	41
22	46

4) Several brands of peanut butter were rated for quality. The following graph compares the price per ounce (in cents) and the quality rating (scale of 0 = lowest to 100 = highest) for each of these brands of peanut butter.



   a) Identify the explanatory and response variables for this situation.

   b) Describe what the scatterplot shows. Be sure to address strength, context, outliers, form and direction (S.C.O.F.D.).

5) Mr. Exercise wanted to know whether or not customers continued to use their equipment after they purchased it. He contacted an SRS of his customers who had purchased an exercise machine during the past 18 months. His findings are summarized in the following table:

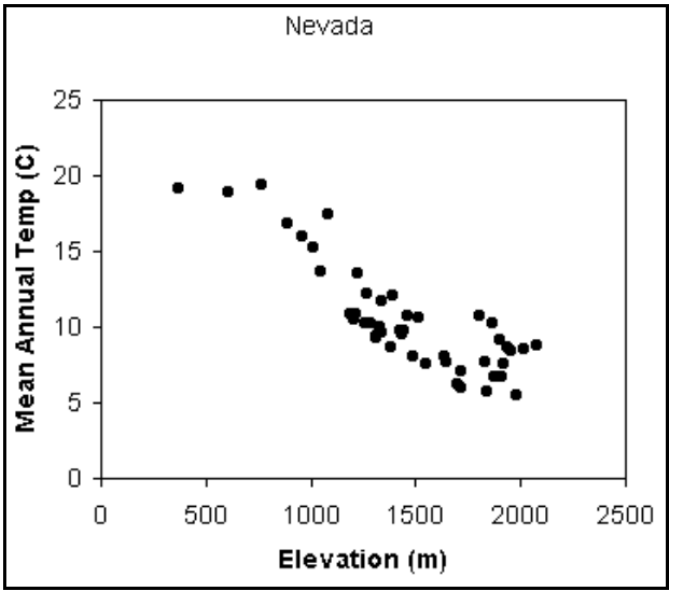
# months owned machine	# hours exercise per week
1	8
5	4.5
7	3
4	6
9	2
14	1.5
5	7
11	4
3	6.5
6	4

   a) Identify the explanatory and response variables for this situation.

   b) Construct a scatterplot to show the results.

   c) Describe what the scatterplot shows. Be sure to address strength, context, outliers, form and direction (S.C.O.F.D.).

6) The following scatterplot shows the elevation and mean temperature for various locations in Nevada.



   a) Identify the explanatory and response variables for this situation.

   b) Describe what the scatterplot shows. Be sure to address strength, context, outliers, form and direction (S.C.O.F.D.).

Review Exercises

7) If two cards are drawn from a standard deck of playing cards and laid face up on a table, what is the probability of getting two Queens?

8) A card is drawn from a standard deck. The card is put back, the deck is reshuffled, and another card is drawn. What is the probability of drawing two clubs?

9) A gum ball machine contains 14 pink, 7 blue, 9 white, and 11 green gumballs. A child buys two gumballs, one after the other. Find the following probabilities:

   a) P(blue, then green)

   b) P(neither is pink)

6.2 Correlation

Learning Objectives

- Understand the properties of the linear correlation coefficient
- Estimate and interpret linear correlation coefficients
- Understand the difference between correlation and causation
- Identify possible lurking variables in bivariate data
- Understand the effects outliers and influential points can have on correlation

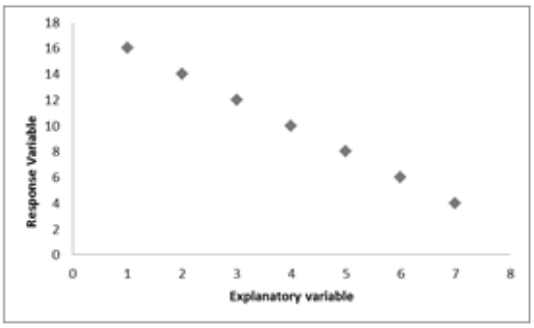
The Correlation Coefficient


https://bit.ly/probstatsSection6-2
(2 videos in this section)

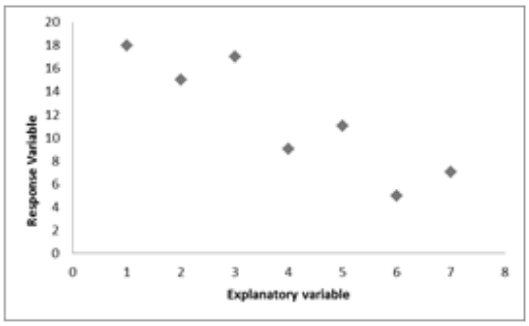
The **correlation coefficient** is a statistic that measures the strength and direction of a *linear* relationship between two numeric variables. The symbol for correlation is **r**, and r can take any value from -1 to +1. The correlation coefficient, r, tells us two things about the linear relationship between two variables, its strength and its direction. The **direction** of the relationship, positive or negative, is given by the sign of the r value. A positive value for r indicates that the relationship is positive (increasing to the right), and a negative r value indicates a negative relationship between the two variables (decreasing to the right). Bivariate data with a positive correlation tells us that as the explanatory variable increases, so does the response variable. Bivariate data with a negative correlation tells us that as the explanatory variable increases, the response variable decreases. A correlation of zero indicates neither of these trends.

The correlation coefficient also tells us about the **strength** of the linear relationship. It tells us how close the points are to forming a perfect line. A correlation (r-value) of exactly 1 or -1 has a perfect correlation. In other words the relationship will produce a scatterplot whose points lie in a perfectly straight line. An r-value of exactly +1 means that the relationship forms a perfect line with a positive slope and an r-value of exactly -1 means that the scatterplot will show a perfect line with a negative slope. The closer the correlation value is to either +1 or -1, the stronger the linear relationship is. As r gets closer to zero (either positive or negative), the weaker the linear relationship is. It is important to note that this is **only measuring the linear relationship** between the two variables. If the relationship shows a clear curved pattern for example, the correlation will tell us nothing about the strength of the relationship.

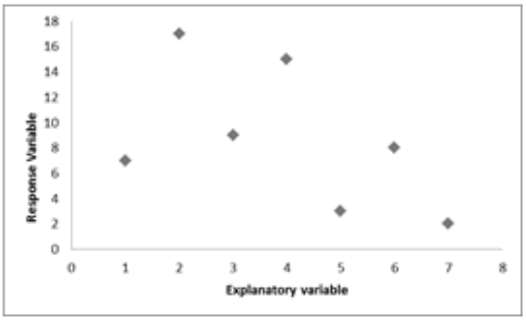
Here are some sample scatterplots with their correlation coefficients given:



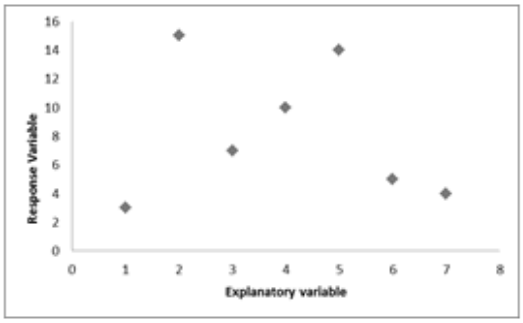
r = -1



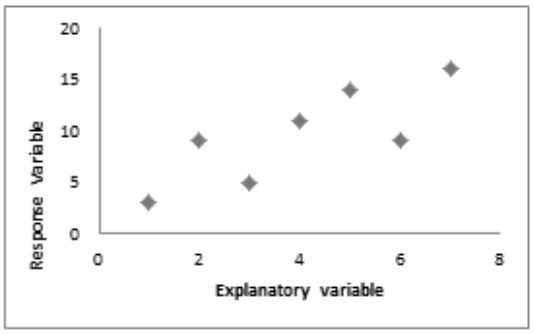
r = -.900



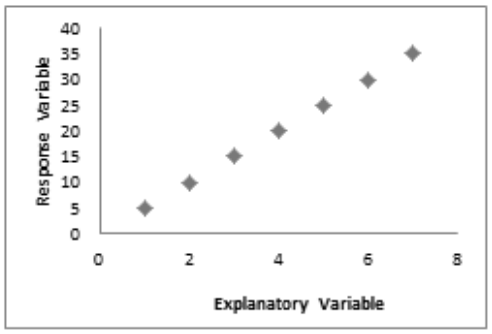
r = -0.536



r = -.160



r = +0.803



r = +1

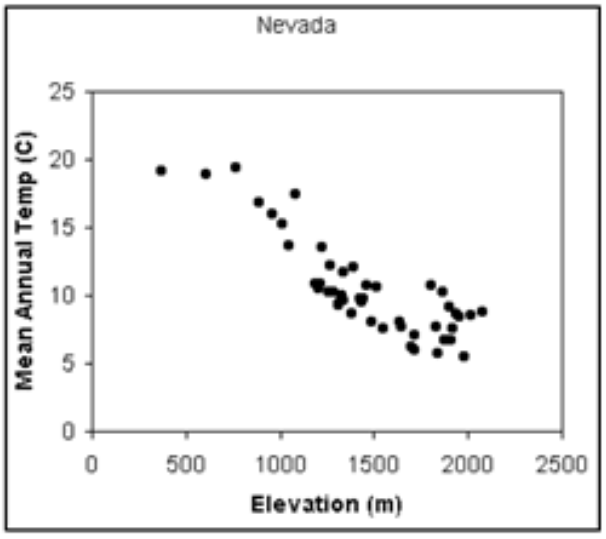
We will be using either our calculator or a computer to **calculate** the correlation coefficient. The formula to calculate the correlation coefficient is quite tedious. It involves calculating the mean and standard deviation of all of the x-values and the mean and standard deviation of all of the y-values. It then compares the x-value of each ordered pair to the mean of x and every y-value to the mean of y (by subtracting and then dividing by the standard deviation), multiplies these newly calculated values, adds all of them, and divides by one less than the sample size. The correlation formula is shown on the next page, but we will be using technology rather than calculating by hand. See Appendix C for calculator instructions.

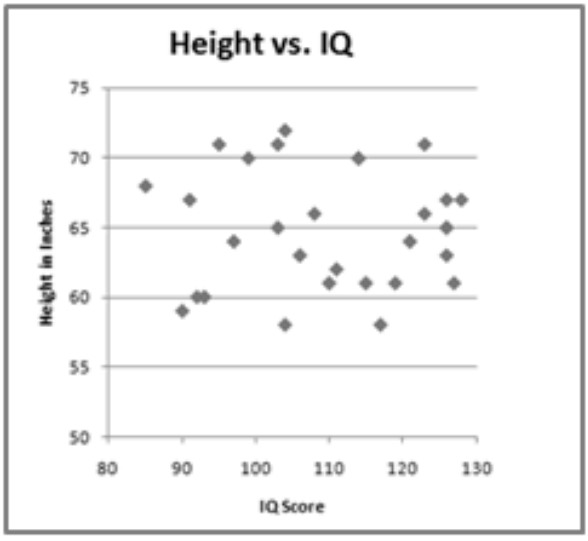
Correlation Coefficient Formula:

$$r = \frac{\sum \left(\frac{(x_i - \bar{x})}{s_x} \right) \left(\frac{(y_i - \bar{y})}{s_y} \right)}{n - 1}$$

Example 1

Estimate the correlation coefficient for each of the following scatterplots.





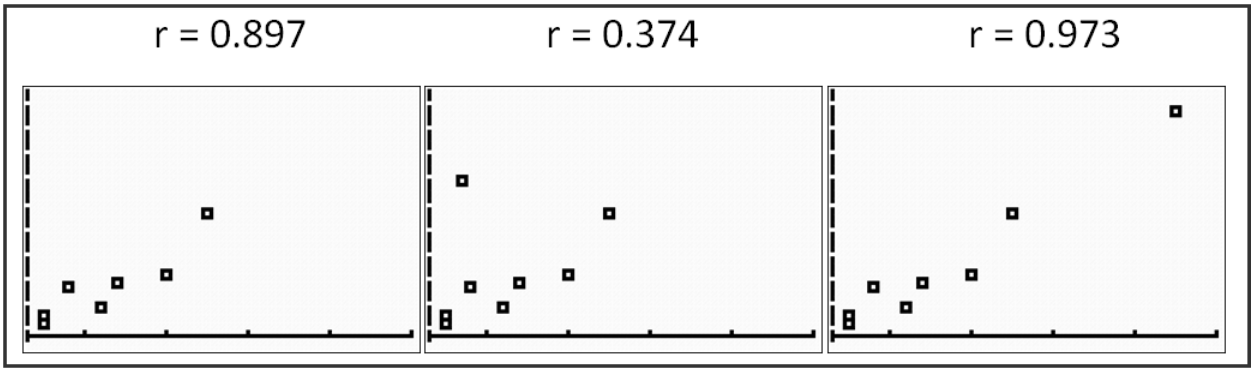
Solution

Nevada: The correlation will be negative and fairly strong, so my estimate is r ≈ -0.85.
Height & IQ: There seems to be no pattern to the graph, so my estimate is r ≈ 0.

Properties of Correlation

When considering using correlation as a measure of the strength between two variables, you should construct and examine a scatterplot first. It is important to check for outliers, be sure that the relationship appears to be linear, be sure that your sample size is sufficient, and consider whether the individuals being examined were too much alike in some way to begin with. Thus, when examining correlation, there are four things that could affect our results: outliers, linearity, size of the sample, and homogeneity of the group.

An outlier, or a data point that lies outside of our overall pattern, can have a significant impact on the correlation. How great of an impact is determined by the sample size of the data set and by how much the outlier lies outside of the pattern of the data. The three plots below show scatterplots with their correlation coefficients (r). The first plot shows a positive and reasonably linear graph. Its correlation is r = 0.897, which is positive and fairly strong. The second plot shows the same data as plot one, with one outlier added in the upper left. Its correlation has dropped to r = 0.374, which is still positive, but much weaker. This demonstrates how outliers can bring the correlation closer to zero. However, some outliers can actually strengthen the correlation. This is demonstrated in the third plot, which shows the same data as the first with one outlier added in the upper right. With this outlier, the linear relationship becomes even stronger than the first plot with r = 0.973.



If the relationship is not linear, calculating the correlation coefficient is meaningless. It is only testing the linear relationship between the two variables. Imagine a scatterplot that shows a perfect parabolic relationship. We would know that there is a strong relationship between these two variables, but if we calculated the correlation coefficient, we would arrive at a figure around zero. Therefore, the correlation coefficient might not always be the best statistic to use to understand the relationship between variables.

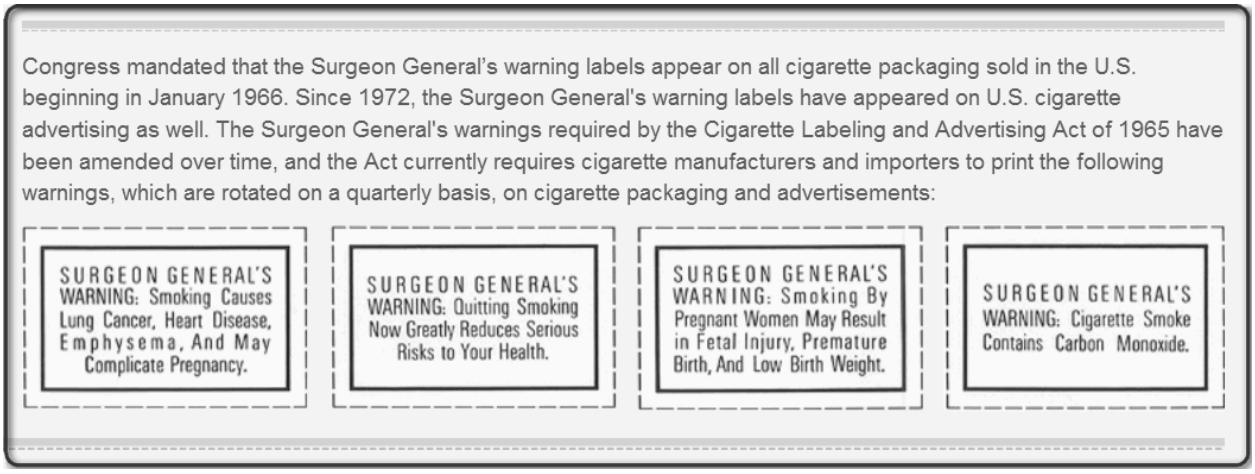
As we discussed in experimental design, a small sample size can be misleading. It can either appear to have a stronger or weaker relationship than is really accurate. The larger the sample, the more accurate of a predictor the correlation coefficient will be on the linearity of two variables.

When a group is too much alike in regard to some characteristics (homogeneous), the range of scores on either or both variables is restricted. For example, suppose we are interested in finding out the correlation between IQ and salary. If only members of the Mensa Club (a club for people with IQ's over 140) are sampled, we will most likely find a very low correlation between IQ and salary since most members will have a consistently high IQ, but their salaries will vary. This does not mean that there is not a relationship between IQ and salary. It simply means that the restriction of the sample limited the magnitude of the correlation coefficient.

Correlation is just a number and it has no units. A change in units of measurement will not impact the correlation. For example, suppose you calculated the correlation coefficient between height in inches and weight in pounds for a group of teenagers. If you later decided to convert the heights to centimeters or the weights to kilograms (or both), and then calculated the correlation coefficient again, you would have found that the value for 'r' did not change.

Lurking Variables

It is very important to know that a high correlation does not mean causation! Oftentimes, studies that show a high correlation between two variables will influence readers into thinking that one variable is the cause of the observed relationship. This is not always true! While in some situations we would agree that one variable is in fact causing the response in the other, it is important to remember that a high correlation simply does not *prove* that one variable is causing the other. The best way to prove such a **direct cause-and-effect** relationship is by carrying out a well-designed experiment. For example, smoking is strongly correlated with lung disease. Based upon much scientific evidence, we can now say that cigarette smoking is one cause of lung disease. However, this topic was highly debated for many years before the surgeon general announced that it was accepted that cigarette smoking causes lung cancer and emphysema. Many people refused to accept this for many years. People who stood to lose money if smoking was proven to be unsafe, suggested many other possible explanations. They suggested that it was simply a coincidence, or that all people who choose to smoke might have something else in common that was actually the cause of the lung disease, not the cigarettes. Because it was not ethical to experiment on humans in order to prove the direct cause-and-effect relationship, the debates went on for a long time.



Sometimes the relationship between variables is cause-and-effect, but many times it can be simply a coincidence that the two variables are highly correlated. It is also possible that some other outside factor, a **lurking variable**, is causing both variables to change. A situation where we have two variables that are both being impacted by some other outside lurking variable is called **common response**. For example, we can show a high correlation between the number of TV's per household and the life expectancy per person among many countries. However, it makes no sense that TV's cause people to live longer. Some lurking variable is playing a role here. It is likely that the economic status of the countries is causing both variables to change; more money means more TV's and more money also means better health care. If a country is wealthy, it is much more likely to have citizens who own TV's. Also, if a country is wealthy, it is much more likely to have good hospitals, roads, health education, and access to food and clean water. These all contribute to a longer life.

In some situations we will have two variables that are highly correlated, but we are unsure of the exact nature of the relationship. We may be unclear as to whether or not one is causing the other, if there is an outside factor impacting the response variable, or if there is some unknown lurking variable that is related in some other unknown way. Remember, lurking variables are not always obvious to the researchers. Such a situation is called **confounding**, because it is confusing to determine how the variables are related (if at all), and whether there may be some lurking variable and if it is related to the variables in question. The variables seem all mixed up and the relationship is unclear, even if highly correlated. An example of confounding is global warming. This is a highly debated topic in social media and web-blogs. Some people argue that human pollution is a major cause of the increase in CO2 and other greenhouse gasses in the atmosphere. Others will argue that it is a part of a natural cycle that has normally occurred in our Earth's history. Still some may think both explanations are at work. This is an example of confounding because there is confusion about the cause of global warming.

Finally, don't forget that some relationships are occurring completely by chance, and their high correlation is then just a **coincidence**. For example, if you researched divorce rates and gas prices over the past 50 years you may note that both have gone up. A scatterplot comparing divorce rates and gas prices would show a strong positive relationship. The correlation would likely be a high, positive value. However, it makes no sense that divorce rates are causing high gas prices. It also is unlikely that there exists a common response or some form of confounding. So in this case, we would say that this is a relationship that is best explained by sheer coincidence.

Example 2

Suggest possible lurking variables to explain the high correlations between the following variables. Explain your reasoning. Consider whether common response, confounding, or coincidence may be involved.

a) It has been shown that cities with more police officers also have higher numbers of violent crimes. Does this mean that more police officers are causing more violent crimes to occur?

b) Over the past 25 years, the percent of parents using car-seats has increased significantly. During this same time period, the rate of DUI arrests has also increased significantly. These two variables, when graphed, show a very high, positive correlation. Does this mean that car-seat use is causing DUI's to increase?

c) A recently published study claimed that, "Teens who use social media a lot [are] more likely to try sex, drugs, and alcohol." Does this mean that social media use causes teens to try sex, drugs and alcohol? Could we then limit teen behaviors such as these by eliminating social media?

Solutions

a) It makes no sense that the number of police officers would be causing the violent crime to occur. It is much more likely that this situation is reversed. Communities with high numbers of violent crimes need higher numbers of police officers. It is also probable that both variables increase in cities with higher populations. Due to the fact that we can think of more than one possible lurking variable and it is difficult to know how all of these variables actually relate, we would say that this is an example of confounding. The variables in question and the lurking variables are all mixed up.

b) It is clearly ridiculous to think that car-seat use is causing an increase in the rate of DUI's. It also makes no sense that DUI's cause car-seats to be used. It could be argued that this is simply a coincidence that these variables are both increasing.

Another possible argument is that there has been in increase in enforcement of laws for both over this time period. The awareness of the dangers of both have increased over the past 25 years, so maybe this is an example of common response.

A third explanation may be that many factors that have contributed to the increase of both, so perhaps this is an example of confounding. Our only certainty is that this is not cause-and-effect. Analysis of these sorts of situations can be very tricky!

c) It is unlikely that social media is actually the cause of these behaviors. There are most likely some lurking variables that are contributing factors. One probable lurking variable, when it comes to teenagers, is the parents. Perhaps this is an example of a common response to parents who are not very involved in their teens' lives. Parents who are not very involved would not be aware that their teen is using social media too much and would also not be aware of what choices their teen is making during his or her free time. Perhaps teens that spend a lot of time unsupervised would be more likely to use social media and would also be more likely to try sex, drugs, and alcohol. All of these behaviors might be a common response to not having parents who prohibit or limit teens from doing these things. Eliminating social media would likely have little to no impact on other teen behaviors.

Multimedia Links:

Calculating Correlation on the Internet,

There are several websites where you can enter in data points and find correlations. You might find the two links below helpful in your understanding of correlation.

http://easycalculation.com/statistics/correlation.php

http://bcs.whfreeman.com/tps5e/default.asp#923932__929340__

Another, more lighthearted example of Correlation ≠ Causation can be found at http://www.exrx.net/ExInfo/Pickles.html which discusses the evil of the pickle.

For a better understanding of correlation try these fun links below,

http://www.istics.net/stat/Correlations

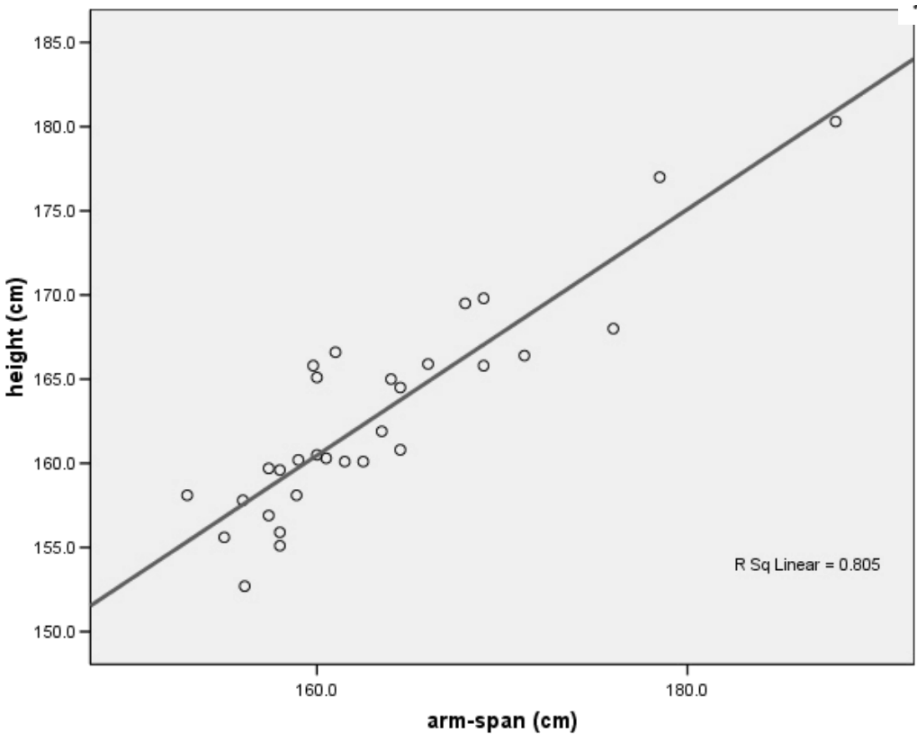
http://www.rossmanchance.com/applets/guesscorrelation/GuessCorrelation.html

Problem Set 6.2

Exercises

1) What are the two things that the correlation coefficient measures?

2) The program used to create this scatterplot found the line-of-best-fit and reported the r-squared value as $r^2 = 0.805$ for the relationship between arm-span and height for several individuals. What is the correlation coefficient? Is it positive or negative? Explain how you know.



3) During the summer Ms. Statsteacher lets her two daughters stay up later than during the school year. Their bedtimes during the summer range from 8:30 p.m. to 12:30 a.m. She has discovered that her older daughter Reily will wake up between 8:00 a.m. and 9:00 a.m. no matter what time she goes to bed. However, her younger daughter Neila tends to wake up later after she gets to stay up later, and earlier when she goes to bed earlier. Neila has been known to wake up anytime between 8:00 a.m. and 11:45 a.m.

   a) Create a separate scatterplot for **each** daughter that compares time going to sleep and time waking up. You will have to approximate your own points. *Which variable will be explanatory and which will be response?*

   b) Which of these do you think will best approximate the correlation for Reily?

      [A] close to r = +1 [B] close to r = +.75 [C] close to r = 0

      [D] close to r = -.75 [E] close to r = -1

   c) Which of these do you think will best approximate the correlation for Neila?

      [A] close to r = +1 [B] close to r = +.75 [C] close to r = 0

      [D] close to r = -.75 [E] close to r = -1

4) Suggest possible lurking variables to explain the high correlations between the following variables. Explain your reasoning. *Consider whether common response, confounding, or coincidence may be involved.*

   a) As ice cream sales increase, the rate of drowning deaths increases sharply. Does this mean that ice cream causes drowning?

   b) With a decrease in the number of pirates, there has been an increase in global warming over the same time period. Does this mean global warming is caused by a lack of pirates?

   c) The higher the number of fire-fighters fighting a fire, the more damage done by the fire. Does this mean that we can limit damage by sending fewer fire-fighters to fires?

   d) Suppose that each of the hockey players on the high school team supplies his or her own hockey stick, with varying degrees of flex. The assistant coach has been keeping a record of the degree of flex for each player's stick and their respective point totals (goals and assists). He has noted that there is a strong, negative correlation between these two variables. In other words, the players with less flex in their sticks are scoring more points and those with more flex are scoring fewer points. Does this prove that the amount of flex in a stick will causes the point totals for the players? Would we be able to give players less flexible sticks and expect to increase scoring?

5) In a recent study in *Resource Manual*, it was noted that divorced men were twice as likely to abuse alcohol as married men. The authors concluded that getting divorced caused alcohol abuse. Do you agree? Explain your reasoning.

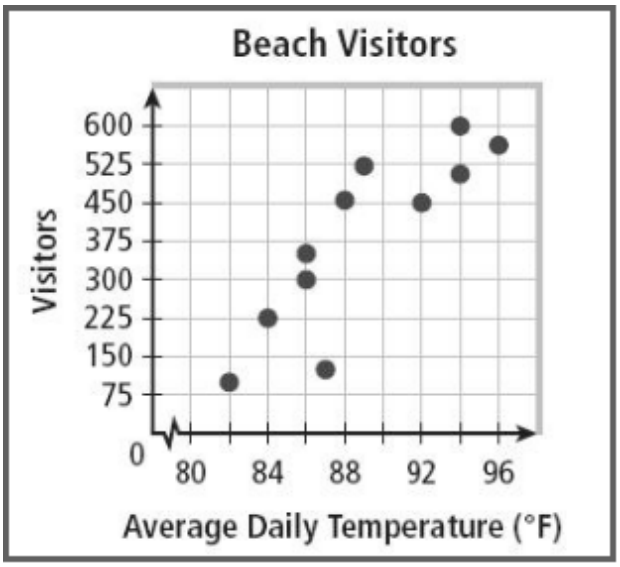
6) A commercial for a new diet pill claims *"You will lose weight while you sleep! No exercise needed!"* They then show several before-and-after photos of people who have lost weight. People who were obese are now very buff. They then give the information for you to order the pills (*"for three payments of just$19.95 each, plus shipping and handling"*). Is this proof that these diet pills caused these people to lose weight? Suggest possible lurking variables. Explain your reasoning.

7) Use the "Beach Visitors" scatterplot to answer the questions that follow.

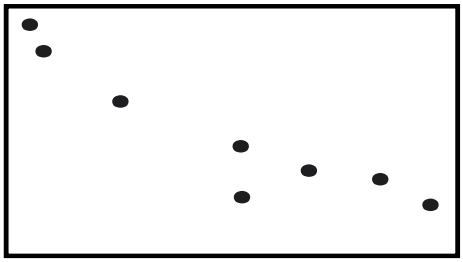
   a) Identify the explanatory and response variables.

   b) Estimate the correlation coefficient for the graph.

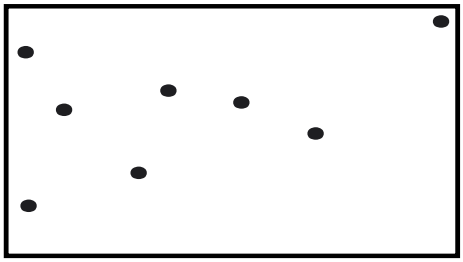
   c) Describe what the scatterplot shows. (*remember S.C.O.F.D.*)



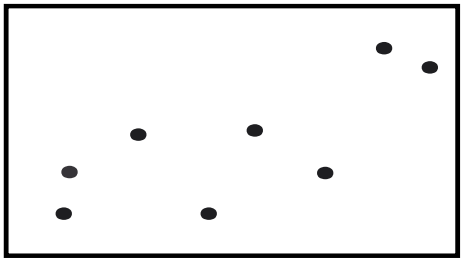
8) Match each graph with its correlation coefficient:



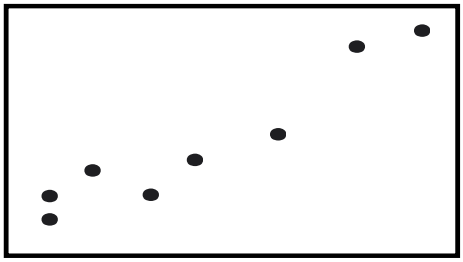
GRAPH #1



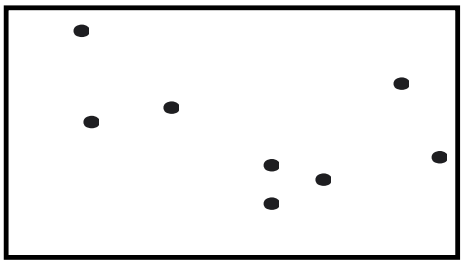
GRAPH #2



GRAPH #3



GRAPH #4



GRAPH #5

Match the correlation with the graph:

A. r = 0.941
B. r = 0.850
C. r = 0.321
D. r = – 0.598
E. r = – 0.938

9) A correlation of r = 0 indicates no linear relationship between the two given variables. But, this does not mean that there is no relationship between the two variables. Sketch a scatterplot in which there is a strong relationship between the variables, but the correlation would be near r = 0.

Review Exercises

10) Zeke flips a coin 93 times and tails shows up 34 of those times. Based on these results, what is the experimental probability of getting tails?

11) If Stephanie's batting average is 0.258, how many hits would you expect her to get out of her next 20 times at bat?

12) You have been playing the game Yahtzee with some friends and you have been keeping track of how often someone gets a Yahtzee (5 of the same dice) when they roll all 5 dice at once. The results today have been 3 Yahtzee's on a single roll, out of 79 trials. Based on these results, what is the experimental probability of getting a Yahtzee in one roll?

13) What is the theoretical probability of getting a Yahtzee in one roll?

6.3 Least-Squares Regression

Learning Objectives

- Construct scatterplots using technology
- Calculate and graph the least-squares regression line using technology
- Calculate the correlation coefficient using technology
- Use the LSRL to make predictions
- Understand interpolation and extrapolation
- Interpret the slope and the y-intercept of the LSRL

https://bit.ly/probstatsSection6-3
(1 video in this section)

Least-Squares Regression

In the last section we learned about the concept of correlation, which we defined as the measure of the strength of a linear relationship between two numerical variables. We saw that when the points of a scatterplot formed a clear linear pattern, we expected a high correlation. Scatterplots can have a strong correlation in either a positive (increasing to the right) or a negative (decreasing to the right) direction. We have also discussed the idea of drawing a line-of-best-fit through the data. In some scatterplots this is easy to do and all of us would end up with our lines in nearly the same place. However, if everyone were to simply draw a line where they thought it fits best, our lines and equations would almost certainly vary from person to person. To maintain consistency, we will use a specific formula to calculate the equation for the line-of-best-fit.

Linear regression involves using data to calculate a line that best fits the data and then using that line to predict scores. We will use the **Least-Squares Regression Line (LSRL)**. The LSRL is the line that makes the sum of the squares of the vertical distance of each data point from the line the least possible value. This is the standard regression equation that is used most often. It is the equation that most calculators and spreadsheets will calculate for you. The formula and process to calculate this is quite tedious, so we will use technology to find the LSRL equation. The regression equation will be in the form $\hat{\mathbf{y}} = \mathbf{a} + \mathbf{bx}$, where **a** is the y-intercept and **b** is the slope of the equation. Your calculator will calculate the correlation coefficient (**r**) at the same time as it calculates the LSRL equation. Many will also report a value for r^2 (which is exactly what it says; r-squared). The $\mathbf{r^2}$ value is called the coefficient of determination. It reports the percent of variation in our data that is explained by our LSRL equation. We will not be addressing its importance in this course.

To calculate the LSRL equation and correlation coefficient, use a graphing calculator, certain scientific calculators, or a computer program. See the Appendix C for instructions on finding the LSRL and correlation coefficient for certain graphing and scientific calculators.

Least-Squares Regression Equation

$$\hat{y} = a + bx$$

x = the explanatory variable

$\hat{y}$ = the predicted response variable

a = the y-intercept (*or the value of y, when x = 0*)

b = the slope (*or the rate of change in y for each increase of one unit in the x direction*)

Interpreting the slope and y-intercept

As with all of our statistics, these data, graphs, and equations are not meaningless. They represent the relationship between two numerical values measured on several specific individuals. Thus the slope and the y-intercept of our newly calculated regression equation mean something as well. We will be interpreting both the slope and y-intercept in context. Our **interpretation of the slope** of the regression equation will be the average rate of change in the response variable ($\hat{y}$) for each increase of one unit of the explanatory variable (x). You will say something like: *For each increase of one (explanatory variable), there will be average (an increase or decrease) of (slope value) in the (response variable).*

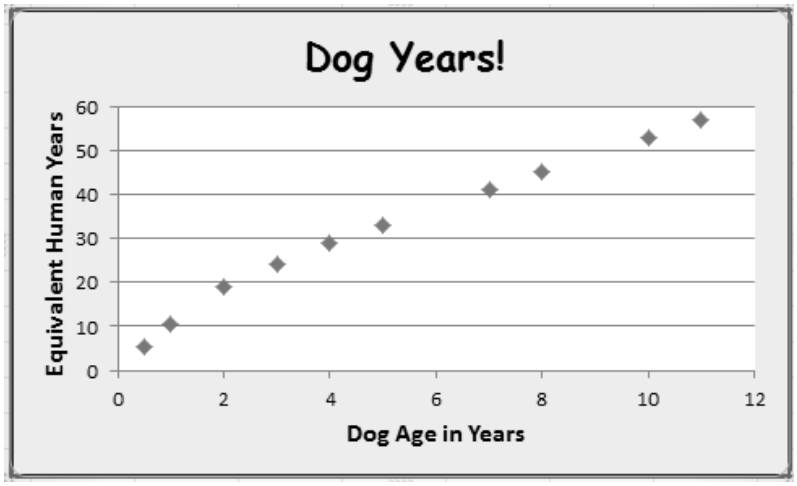
Our **interpretation of the y-intercept** of the regression equation will be the predicted value of the response variable (y) when the explanatory variable (x) is zero. You will say something like: *When (explanatory variable) is zero, the (response variable) is predicted to be (y-intercept value)*. You will discover that the interpretation of the y-intercept often makes absolutely no sense when put into context. This is because actual data often does not involves x-values of zero.

Example 1

To the right is data given by a canine expert. It relates a dog's age in years to what they believe the equivalent age in human years to be.

Dog Age (in Years)	Equivalent Human Age (in Years)
0.5	5.5
1	10.5
2	19
3	24
4	29
5	33
7	41
8	45
10	53
11	57

The scatterplot showing this data, using dog age as the explanatory variable, is shown below.



a) Calculate the Least-Squares regression line for the Dog Year Data. Report your equation. Be sure to identify your variables.

b) Calculate the correlation coefficient (r). What two things does r tell us about this relationship?

c) Identify and interpret the slope in the context of the problem.

d) Identify and interpret the y-intercept in the context of the problem.

Solution

a) $\hat{y} = 7.7947 + 4.6418x$ where x is the age of the dog in years and $\hat{y}$ is the equivalent predicted age in human years.

b) r = +0.9907. This r-value tells us that the graph is increasing left to right and that this data produces a strong linear relationship.

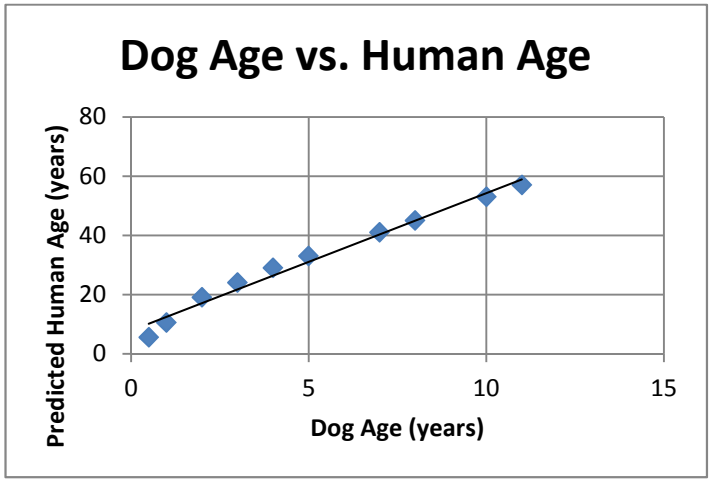
c) The slope is 4.642. This means that for every increase of one year in dog age, there is an average increase of 4.642 years in the equivalent human age.

d) The y-intercept is 7.795. It means that if a dog were 0 years old, it would be predicted to be 7.795 years in human years. (Note that this clearly makes no sense in this case. It is reasonable that both values would start at zero.)

Making Predictions

The main use of the regression line is to predict values. After calculating this line, we are able to predict values by simply substituting a value for the explanatory variable (x) and solving the equation for the predicted response value ($\hat{y}$). In our example above, we can predict that the human year equivalence for a dog that is 6 years old is approximately 35.6 human years.

$\hat{y} = 7.795 + 4.642(6) = 35.647$. This prediction is reasonable and it matches our graph. However this is not always the case.



As you look at the LSRL drawn on the above scatterplot, you can see that the points to the far left do not appear to be very linear. So, using the line to the left of about 1 year will not make much sense. Also, we do not have any idea what will happen to the data beyond the 11 years that we have recorded. An LSRL is very useful in making predictions, but only within the range of the actual data that we have collected and can see. This is called **interpolation**. We can see that this line is a reasonably good fit between 1 and 11 dog years, but we simply do not know what happens beyond 11 years (and we cannot use negative years for obvious reasons). The prediction line that we have calculated will go forever in both directions, but it will not be appropriate to use it to predict for all values of x. Using a regression line to predict values that are outside the range of our actual data is called **extrapolation**. Extrapolation will sometimes yield ridiculous answers! However, even if the result seems reasonable, we should avoid extrapolating because we simply do not know what happens beyond our actual observations. Making decisions based on extrapolating can be dangerous as we are coming to conclusions that are not backed up by data.

Example 2

The table to the right lists the GPA and Verbal SAT Score for seven students. Analyze how well Verbal SAT Scores can be used to predict a student's GPA based on this data.

Student	Verbal SAT Score	GPA
Anna	595	3.4
Bryce	520	3.2
Corbin	715	3.9
Delia	405	2.3
Emilio	680	3.9
Frankie	490	2.5
Geraldine	565	3.5

a) Construct a scatterplot on your graphing calculator (or computer). Sketch the graph that the calculator shows. Be sure to label your axes.

b) Calculate the Least-Squares Regression Line (LSRL) using your calculator and include it on the graph from part a). Be sure to identify your variables.

c) Calculate the correlation coefficient (r). What are the two things that this number tells us about the graph?

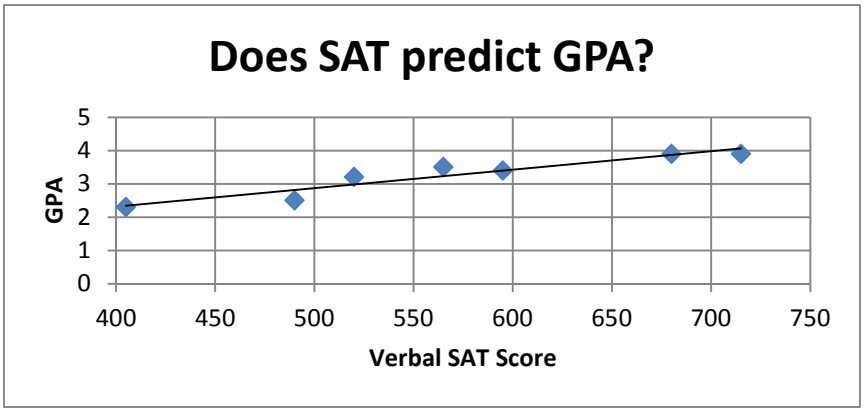
d) Identify and interpret the slope in the context of the problem.

e) Using your equation, what is the predicted GPA of a student who has a verbal SAT score of 500?

f) Using your equation, what is the predicted verbal SAT score of a student who has a GPA of 3.1?

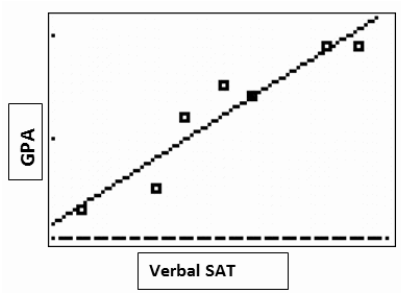
Solution

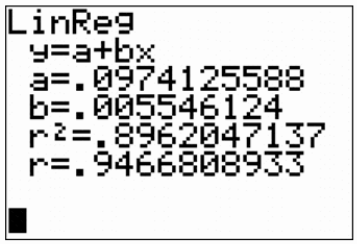
a)



b) $\hat{y} = 0.097 + 0.0055x$ where x stands for the student's verbal SAT score and $\hat{y}$ stands for the student's GPA.

Here are the LSRL, correlation, and the scatterplot with the line added to the graph, from a screen shot of a TI-84 plus:





c) The correlation is r = +0.9467. This tells us that the relationship is positive and strong.

d) The slope is 0.0055. This tells us that for each increase of 1 point on the verbal SAT score, there will be an average increase of 0.0055 points in a student's GPA.

e) $\hat{y}$ = 0.097 + 0.0055(500) = 2.847 The predicted GPA for a student who scores 500 on the verbal portion of the SAT exam is 2.847.

f) 3.1 = 0.097 + 0.0055x

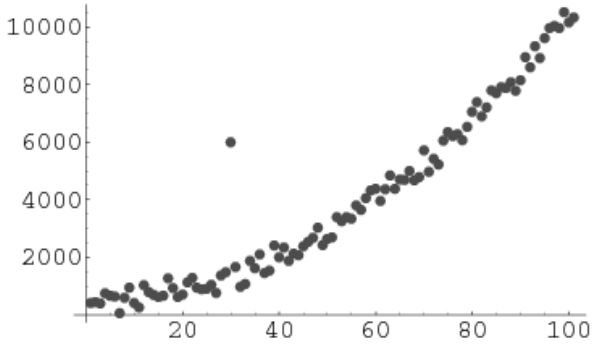
3.003 = 0.0055x

546 = x

The predicted SAT verbal score for a student with a GPA of 3.1 is 546.

Outliers and Influential Points

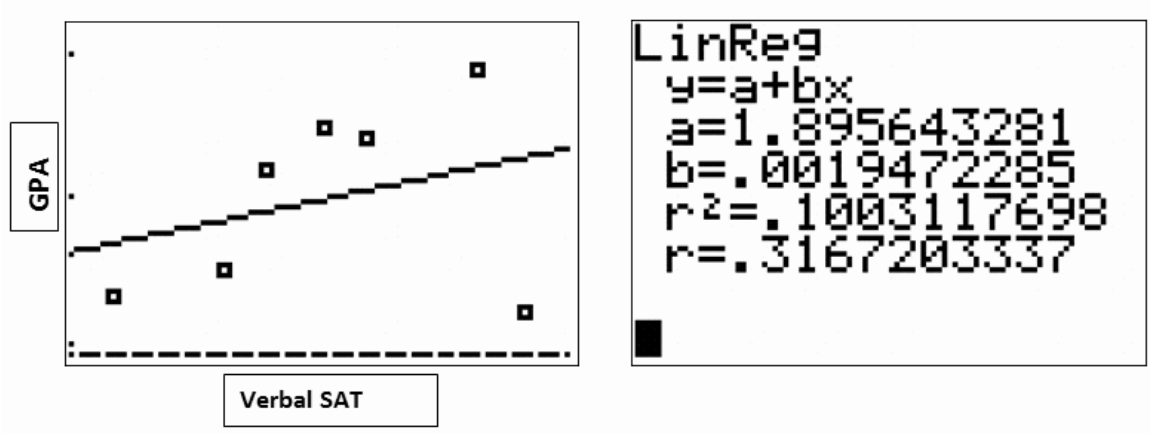
An outlier is an extreme observation that does not fit the general pattern of the data (see the example data set to the right). Because an outlier is an extreme observation, the inclusion of it may impact both the correlation and the equation for the least-squares regression line. When examining a scatterplot and calculating the regression equation, it is worth considering whether extreme observations should be included or not.



Let's use our GPA example to illustrate the effect of a single outlier. Suppose that we have a student who has scored very high on the SAT Verbal exam, but has a lower GPA. We will change Corbin's GPA from 3.9 to 2.2 and see what happens to the LSRL and correlation.

Student	Verbal SAT Score	GPA
Anna	595	3.4
Bryce	520	3.2
Corbin	715	2.2
Delia	405	2.3
Emilio	680	3.9
Frankie	490	2.5
Geraldine	565	3.5

Here is screen shot of aTI-84 with the adjusted score for Corbin. Both the LSRL and the correlation changed.



As you can see, this one change turned Corbin into an outlier. This caused the correlation to drop from r= +0.947, all the way down to r = +0.317. This one data point caused a huge change. It made the relationship between the two variables extremely weak rather than very strong. In addition, the adjusted data point changed both the slope and the y-intercept of the LSRL equation dramatically. This means that predictions based on this LSRL will have different results than those based on the LSRL with Corbin's old GPA.

There is no set rule when trying to decide how to deal with outliers in regression analysis, but you can now see how an outlier can dramatically change everything when it comes to scatterplots, correlation, and least-squares regression equations. Be sure to mention any potential outliers that you observe in any scatterplot.

Multimedia Links

For an introduction to what a least squares regression line represents,

See Bionic Turtle at http://www.youtube.com/watch?v=ocGEhiLwDVc (5:15).

For an applet that will calculate correlation and the least squares regression line,

Visit http://illuminations.nctm.org/lessonDetail.aspx?ID=L456

Problem Set 6.3

Exercises

1) Malia turned the water on in her bathtub full blast. She then measured the depth of the water every two minutes until the bathtub was full. Her findings are listed in the following table. In section 6.1 we constructed a scatterplot and described the plot, we are now going to analyze this data further.

Time (minutes)	Depth (cm)
2	7
4	9.5
6	14
8	19.5
10	21
12	24
14	32
16	36
18	37.5
20	41
22	46

   a) Construct a scatterplot on your graphing calculator (or computer). Sketch the graph that the calculator shows. Be sure to label your axes. Use a ruler to draw in a best-fit line.

   b) Calculate the Least-Squares Regression Line (LSRL) using your calculator. Report your equation. Be sure to identify your variables.

   c) Calculate the correlation coefficient (r). What are the two things that this number tells us about this graph?

   d) Identify and interpret the slope in the context of the problem.

   e) Using your equation, what is the predicted depth of the water after 17 minutes? After one hour? Are these answers reasonable? Why or why not?

2) The following table shows the progression of the Federal Minimum Wage in the United States since 1938 (source:http://www.laborlawcenter.com). We are going to analyze the relationship between year and minimum wage to see if there is a predictable relationship between the variables.

FEDERAL MINIMUM WAGE HISTORY

Effective Date	*Hourly Wage*
10/24/1938	$0.25
10/24/1939	$0.30
10/24/1945	$0.40
01/25/1950	$0.75
03/01/1956	$1.00
09/03/1961	$1.15
09/03/1963	$1.25
02/01/1967	$1.40
02/01/1968	$1.60
05/01/1974	$2.00
01/01/1975	$2.10
01/01/1976	$2.30
01/01/1978	$2.65
01/01/1979	$2.90
01/01/1980	$3.10
01/01/1981	$3.35
04/01/1990	$3.80
04/01/1991	$4.25
10/01/1996	$4.75
09/01/1997	$5.15
07/24/2007	$5.85
07/24/2008	$6.55
07/24/2009	$7.25

   a) Using **year only** as the explanatory variable (ignore month & day), construct a scatterplot. Sketch the graph that the calculator shows. Be sure to label your axes.

   b) Describe the relationship between the two variables. (S.C.O.F.D.)

   c) Calculate the Least-Squares Regression Line (LSRL). Add the line to your graph and report your equation. Be sure to identify your variables.

   d) Calculate the correlation (r). Even though r is very high, do you feel that a line is the best model for this data? Why or why not?

   e) Based on the linear model from part d), what would you predict the Federal Minimum Wage to be in 2016? Is this an accurate prediction? Why or why not?

   f) Based on your model from part d), what would you predict the minimum wage to have been in 1968? How close is this to the actual minimum wage that year?

3) Suppose that some researchers analyzed the relationship between fathers' and sons' IQ scores for a group of men. Suppose they discovered that the relationship was reasonably linear and they calculated a regression line of $\hat{y} = 12 + 0.9x$ where x = father's IQ and $\hat{y}$ = son's predicted IQ.

   a) Identify the explanatory and response variables.

   b) Identify and interpret the slope in the context of the problem.

   c) Identify and interpret the y-intercept in the context of the problem.

   d) Do your answers to (b) and (c) seem reasonable? Why or why not?

   e) What would you predict a son's IQ to be if his father has an IQ of 120? What if the father had an IQ of 140?

   f) If you knew that the original data included fathers with IQs from 108 to 145, explain why it would be inappropriate to use your model to predict a son's IQ if his father's IQ were 170.

4) Mr. Exercise wanted to know whether or not customers continued to use their equipment after they purchased it. He contacted an SRS of his customers who had purchased an exercise machine during the past 18 months. His findings are summarized in the following table. We began to look at this data in section 6.1. We are now going to analyze it further.

   a) Construct a scatterplot. Calculate the LSRL and add it to your graph. Sketch your graph and report your equation. Be sure to identify your variables.

   b) Identify and interpret the slope in the context of the problem.

   c) Identify and interpret the y-intercept in the context of the problem.

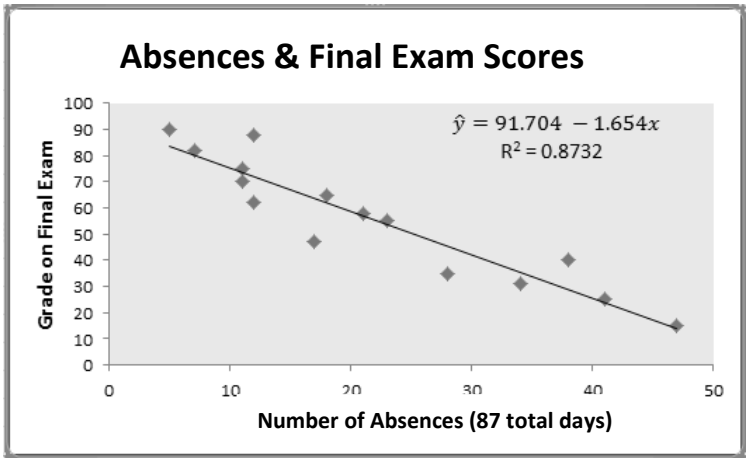
   d) What is the correlation coefficient? What are the two things that this statistic tells about the relationship between these two variables?

   e) Based on your model, how many hours would you predict a person who has owned their machine for 12 months to exercise?

   f) Based on your model, if a person claims to exercise 9 hours per week, how long would you suspect that they had owned the machine?

# months owned machine	# hours exercise per week
1	8
5	4.5
7	3
4	6
9	2
14	1.5
5	7
11	4
3	6.5
6	4

5) A college professor was becoming annoyed by how many of his students were absent during his 8:00 a.m. section of Philosophy 103. He decided to analyze whether these absences were impacting student scores. He assigned his TA the task of keeping track of attendance. At the end of the semester he compared each student's grade on the final exam (100 points possible) with the number of times he or she had been absent. His findings are displayed in the graph to the right.



a) Identify the explanatory and response variables.

b) Describe the relationship between these two variables (S.C.O.F.D).

c) Jeremy was absent 25 times. What would you predict his score on the final exam to be?

d) Lucy often overslept and missed 43 class sessions. What would you predict for her score on the final?

e) Calculate the correlation coefficient (r). What two things does this statistic tell you about the association between these two variables? (Hint: You were given the value for R^2.)

f) Interpret the meaning of -1.654 in the context of this problem.

6) The table to the right shows the grade level and reading level for 5 students. Treat grade level as the explanatory variable as you answer the questions below.

Grade vs. Reading

Student	Grade	Reading Level
A	2	7
B	6	14
C	5	12
D	4	9
E	1	4

a) Create a scatterplot and then calculate the LSRL and the correlation coefficient for this data set.

Suppose it was determined that student E was actually in grade 8. Let's examine how this change would this impact the LSRL and the correlation.

b) Create a scatterplot for the new data. Then calculate the LSRL and the correlation coefficient for the changed data. The new data is shown in the table to the right.

c) What changes do you notice between your answers to (a) and (b)? Explain why these changes occurred.

Grade vs. Reading

Student	Grade	Reading Level
A	2	7
B	6	14
C	5	12
D	4	9
E	8	4

7) The table below gives the nutritional information for Taco Bell Burritos as reported on the website: http://www.tacobell.com. Choose two of the variables to analyze. (Do not use trans fat.)

   a) What will you be using as your explanatory and response variables?

   b) Construct a scatterplot. Label your axes.

   c) Describe the association (S.C.O.F.D.).

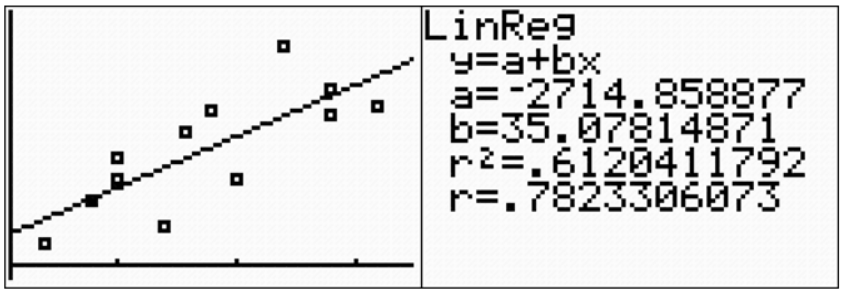
   d) Calculate the LSRL and the correlation. Report them. Be sure to define your variables. Add the line to your graph in part (b).

   e) Use your model to make a prediction that involves interpolation.

   f) Use your model to make a prediction that involves extrapolation.

item	serving size (g)	calories	calories from fat	saturated fat (g)	total fat (g)	trans fat (g)	cholesterol (mg)	sodium (mg)	carbohydrates (g)	dietary fiber (g)
Burritos										
1/2 lb.* Cheesy Potato Burrito	248	540	230	7	26	0.5	45	1360	59	7
1/2 lb.* Combo Burrito	241	460	160	7	18	0.5	45	1330	53	9
7-Layer Burrito	283	500	160	6	18	0	20	1090	69	12
Bean Burrito	198	370	90	3.5	10	0	5	980	56	10
Beefy 5-Layer Burrito	245	540	190	8	22	0	35	1280	68	9
Beefy Nacho Burrito	186	470	180	6	20	0	30	990	58	4
Burrito Supreme® - Chicken	248	400	110	5	12	0	40	1060	51	7
Burrito Supreme® - Steak	248	390	110	5	13	0	30	1100	51	7
Burrito Supreme® – Beef	248	420	140	6	16	0	35	1100	53	9
Chili Cheese Burrito	156	380	150	8	17	0.5	35	930	41	5
Fresco Bean Burrito	213	350	70	2.5	8	0	0	990	57	11
Grilled Chicken Burrito	177	430	170	5	18	0	35	870	48	3
XXL Grilled Stuft Burrito - Beef	445	880	370	14	42	1	75	2050	95	14
XXL Grilled Stuft Burrito - Chicken	445	840	310	11	35	0	85	1970	92	11
XXL Grilled Stuft Burrito - Steak	445	820	320	12	36	0.5	70	2050	92	11

8) Use the calculator output from a screenshot of a TI-84 to answer the questions. A lifeguard at Swimtastic Pool & Water-Slides decided to keep track of how many people came to the pool each day and compare this to the high temperature for that day. The temperatures ranged from 82° to 96° during his data collection time period. He used the number of people as the response variable.



   a) Write the regression equation. Define your variables.
   b) Identify and interpret the slope in the context of the problem.
   c) Report the correlation. What two things does the correlation tells us in this situation?
   d) Based on this model, how many people would you predict on a 91° day? How many would you predict on a 45° day? Are both of these predictions reasonable? Why or why not?

Review Exercises

Use the information below to answer review exercises 9 – 12.

Suppose that Marco, the star of the basketball team, makes 79% of the free-throws that he attempts. Assume that each free-throw is independent of any other free-throw.

9) What is the probability that Marco will make three free-throws in a row?

10) What is the probability that Marco will make exactly two out of three free-throws?

11) What is the probability that Marco will miss at least one of his next four free-throws?

12) If you were going to set up a simulation to estimate this scenario, which of the following would not be an appropriate way to assign the digits?

   [A] 01-79 represents makes, 80-99 & 00 represents misses
   [B] 01-21 represents misses, 22-99 & 00 represents makes
   [C] 00-79 represents makes, 80-99 represents misses
   [D] 00-20 represents misses, 21-99 represents makes
   [E] 00-78 represents makes, 79-99 represents misses

13) Here are the hourly salaries for the employees at Greezy's Burger Boy:
   $9.35, $9.85, $9.25, $10.90, $10.25, $9.25, $12.05, $9.70, $18.90, $10.30, $9.75, and $9.55
   Use this salary data to answer the questions below.
   a) Calculate the mean and standard deviation for the salaries.
   b) Calculate the five number summary for the salaries.
   c) Construct an accurate box plot.
   d) Which numerical measures of center and spread (mean & standard deviation or median & IQR) would be more appropriate in this situation? Explain why.
   e) Describe the distribution. Include Shape, Outliers, Context, Center, & Spread (S.O.C.C.S.)

6.4 Chapter 6 Review

In this chapter, we have learned that when working with bivariate data, it is important to first identify whether there is an explanatory and response relationship between the two variables. Often one of the variables, the explanatory (independent) variable, can be identified as having an impact on the value of the other variable, the response (dependent) variable. The explanatory variable should be placed on the horizontal axis, and the response variable should be placed on the vertical axis. Next we learned how to construct a visual representation, in the form of a scatterplot, so that we can see what the association looks like. A scatterplot helps us see what, if any, association there is between the two variables. If there is an association between the two variables, it can be identified as being strong if the points form a very distinct form or pattern, or weak if the points appear to be somewhat randomly scattered. If the values of the response variable generally increase as the values of the explanatory variable increase, the data has a positive association. If the response variable generally decreases as the explanatory variable increases, the data has a negative association. We also are able to see the form of the pattern, if any, in the graph.

When the data looked reasonably linear, we learned how to use technology to calculate the least-squares regression line and the correlation coefficient. The least-squares regression line is often useful for making predictions for linear data. However, we now know to beware of extrapolating beyond the range of our actual data. Correlation is a measure of the linear relationship between two variables – it does not necessarily state that one variable is caused by another. For example, a third variable or a combination of other factors may be causing the two correlated variables to relate as they do. We learned how to interpret the linear correlation coefficient and that it can be greatly affected by outliers and influential points. Also, just because two variables have a high correlation, does not mean that they have a cause-and-effect relationship. Correlation does not necessarily imply causation!

Beyond constructing graphs and calculating statistics, we learned how to describe the relationship between the two variables in context. The acronym we learned to help us remember what to include in our descriptions is S.C.O.F.D. This tells us to describe the strength of the association, to be sure that our description is in context, to mention any outliers or influential points that we observe, and to describe the form and the direction of the relationship. We also learned how to interpret the slope and y-intercept of the least-squares regression line in context. Even though we are doing easy calculations, statistics is never about meaningless arithmetic and we should always be thinking about what a particular statistical measure means in the real context of the data.

Review Exercises

Answer the following as TRUE or FALSE.

1) A negative relationship between two variables means that for the most part, as the x variable increases, the y variable increases.
2) A correlation of -1 implies a perfect linear relationship between the variables.
3) The equation of the regression line used in statistics is $\hat{y} = a + bx$
4) When the correlation is high, one can assume that x causes y.

Complete the following statements with the best answer.

5) The variable used for the correlation coefficient is _____ .
6) A statistical graph of two variables is called a(n) _____ .
7) The_____ variable is plotted along the x-axis.
8) The range of r is from _____ to _____ .
9) The sign of r and _____ will always be the same.
10) LSRL stands for_____ .
11) If all the points fall on a straight line, the value of r will be _____ or _____ .
12) If r = -0.86, then r^2 = _____ .
13) If r^2 = 0.77, then r = _____ or _____ .
14) Using an LSRL to make predictions outside the range of our original data is called _____ .
15) Using an LSRL to make predictions within the range of our original data is called _____ .
16) When describing the relationship visible in a scatterplot, the acronym S.C.O.F.D. stands for _____ .
17) Suppose that a scatterplot shows a strong, linear, positive relationship, and the correlation coefficient is very high. However, both of the variables are actually increasing due to some outside lurking variable. This relationship is an example of _____ .
18) Suggest possible lurking variables to explain the high correlations between the following variables. Consider whether common response, confounding, or coincidence may be involved.
    a) The number of cell phones being made has been increasing over the past 15 years. So has the number of starving children. Do cell phones cause starvation?
    b) The stress level of all of the employees at a certain company has been going up consistently over the past year. During this time, they have received three pay bumps. Does this mean that higher pay is causing the stress?
    c) Suppose that a study shows that the number of hours of sleep a person gets is negatively correlated with the number of cigarettes a person smokes. In other words, as the number of hours of sleep goes down, the number of cigarettes smoked goes up. Does this mean that not sleeping causes a person to smoke more cigarettes?

19) Some researchers wanted to determine how well the number of beers consumed can predict what a person's blood alcohol content will be after a given length of time. They set up an experiment in which several volunteers each drank a randomly selected number of beers during a given time period. The volunteers were between 21 and 25 years of age, but all ranged in gender and in weight. Exactly three hours after they began to drink the beers, their BAC level was measured three times. The three measurements were averaged and the results are given in the following table. (This is fictitious data, but it is based upon calculations from the BAC calculator at http://www.dot.wisconsin.gov.)

Number of Beers Consumed (3 hours)	10	2	4	6	8	3	3	7	8	5	9	4	6	2	5
BAC Level	0.29	0.034	0.094	0.1	0.135	0.025	0.062	0.23	0.225	0.127	0.137	0.13	0.06	0.012	0.139

a) Identify the explanatory and response variables and construct a scatter-plot (be neat & label your axes).

b) Calculate the LSRL and correlation. Report the equation and add it to your scatter-plot. Identify what your variables represent.

c) Identify and interpret the slope in context.

d) Identify and interpret the y-intercept in context.

e) If a person drinks 6 beers during this time period, on average what do you predict the person's BAC will be?

f) If a person drinks 15 beers during this time period, on average what do you predict the person's BAC will be?

g) Are you confident in both of the previous answers? Why or why not?

h) How many beers would predict had been consumed if the BAC was measured at 0.122?

20) When investigating car crashes, it is often necessary to try to determine the speed at which a vehicle was traveling at the time of the accident. Investigators are able to do this by measuring the length of the skid mark left by the vehicle in question. The table below lists several speeds (mph) based on the skid length (feet), according to the Forensic Dynamics website: http://forensicdynamics.com.

a) Identify the explanatory and response variables and construct a scatterplot. Be sure to properly scale and label your graph.

b) Calculate the LSRL and add it to your scatterplot. Report your equation and identify your variables.

c) Describe the relationship you see in the scatter-plot (S.C.O.F.D.). Be thorough & use complete sentences! Be sure that you explain the relationship in the context of the problem.

d) What is the correlation coefficient? Based on your scatterplot and the value of r, how well do you feel that your model fits this data? Explain.

e) What is the predicted speed if the skid mark is 157 feet?

f) What is the predicted speed for a skid mark of 36 feet?

g) Would you expect predictions beyond 250 feet to generally over-estimate or under-estimate the actual speed of the vehicle? Why?

SPEED BASED ON SKID LENGTH

Skid Length (feet)	Estimated Speed (mph)
45	30.68
20	20.45
56	34.23
8	12.93
78	40.4
93	44.11
165	58.75
115	49.05
142	54.51
184	62.05
215	67.07
247	71.89

Image References:

Beach visitors & temperature: http://technomaths.edublogs.org

Study Time & Test Scores: http://www.icoachmath.com

Car weight & mpg: http://www.statcrunch.com

Elevation & Temperature: http://staff.argyll.epsb.ca

Peanut Butter & Quality Rating: http://intermath.coe.uga.edu

Arm Span & Height: http://3.bp.blogspot.com

Surgeon General's Warning Labels: http://abibrands.com

Outlier Example: http://mathworld.wolfram.com

Recycling Rates: http://www.earth-policy.org

Appendices

Appendix A – Tables

Appendix A, Part 1 - Random Digit Table

Line 101	19223	95034	05756	28713	96409	12531	42544	82853
Line 102	73676	47150	99400	01927	27754	42648	82425	36290
Line 103	45467	71709	77558	00095	32863	29485	82226	90056
Line 104	52711	38889	93074	60227	40011	85848	48767	52573
Line 105	95592	94007	69971	91481	60779	53791	17297	59335
Line 106	68417	35013	15529	72765	85089	57067	50211	47487
Line 107	82739	57890	20807	47511	81676	55300	94383	14893
Line 108	60940	72024	17868	24943	61790	90656	87964	18883
Line 109	36009	19365	15412	39638	85453	46816	83485	41979
Line 110	38448	48789	18338	24697	39364	42006	76688	08708
Line 111	81486	69487	60513	09297	00412	71238	27649	39950
Line 112	59636	88804	04634	71197	19352	73089	84898	45785
Line 113	62568	70206	40325	03699	71080	22553	11486	11776
Line 114	45149	32992	75730	66280	03819	56202	02938	70915
Line 115	61041	77684	94322	24709	73698	14526	31893	32592
Line 116	14459	26056	31424	80371	65103	62253	50490	61181
Line 117	38167	98532	62183	70632	23417	26185	41448	75532
Line 118	73190	32533	04470	29669	84407	90785	65956	86382
Line 119	95857	07118	87664	92099	58806	66979	98624	84826
Line 120	35476	55972	39421	65850	04266	35435	43742	11937

Line 121	71487	09984	29077	14863	61683	47052	62224	51025
Line 122	13873	81598	95052	90908	73592	75186	87136	95761
Line 123	54580	81507	27102	56027	55892	33063	41842	81868
Line 124	71035	09001	43367	49497	72719	96758	27611	91596
Line 125	96746	12149	37823	71868	18442	35119	62103	39244
Line 126	96927	19931	36089	74192	77567	88741	48409	41903
Line 127	43909	99477	25330	64359	40085	16925	85117	36071
Line 128	15689	14227	06565	14374	13352	49367	81982	87209
Line 129	36759	58984	68288	22913	18638	54303	00795	08727
Line 130	69051	64817	87174	09517	84534	06489	87201	97245
Line 131	05007	16632	81194	14873	04197	85576	45195	96565
Line 132	68732	55259	84292	08796	43165	93739	31685	97150
Line 133	45740	41807	65561	33302	07051	93623	18132	09547
Line 134	27816	78416	18329	21337	35213	37741	04312	68508
Line 135	66925	55658	39100	78458	11206	19876	87151	31260
Line 136	08421	44753	77377	28744	75592	08563	79140	92454
Line 137	53645	66812	61421	47836	12609	15373	98481	14592
Line 138	66831	68908	40772	21558	47781	33586	79177	06928
Line 139	55588	99404	70708	41098	43563	56934	48394	51719
Line 140	12975	13258	13048	45144	72321	81940	00360	02428
Line 141	96767	35964	23822	96012	94591	65194	50842	53372
Line 142	72829	50232	97892	63408	77919	44575	24870	04178
Line 143	88565	42628	17797	49376	61762	16953	88604	12724
Line 144	62964	88145	83083	69453	46109	59505	69680	00900
Line 145	19687	12633	57857	95806	09931	02150	43163	58636
Line 146	37609	59057	66967	83401	60705	02384	90597	93600
Line 147	54973	86278	88737	74351	47500	84552	19909	67181
Line 148	00694	05977	19664	65441	20903	62371	22725	53340
Line 149	71546	05233	53946	68743	72460	27601	45403	88692
Line 150	07511	88915	41267	16853	84569	79367	32337	03316

Appendix A, Part 2 – The Normal Distribution Table

For z-scores with z less than or equal to zero

Table 8.1

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3829
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

For z-scores with z greater than or equal to 0

Table 8.2

z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0..6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Appendix A, Part 3 – A standard deck of 52 cards

Black cards		*Red cards*	
Clubs	**Spades**	**Hearts**	**Diamonds**
A ♣	A ♠	A ♥	A ♦
2 ♣	2 ♠	2 ♥	2 ♦
3 ♣	3 ♠	3 ♥	3 ♦
4 ♣	4 ♠	4 ♥	4 ♦
5 ♣	5 ♠	5 ♥	5 ♦
6 ♣	6 ♠	6 ♥	6 ♦
7 ♣	7 ♠	7 ♥	7 ♦
8 ♣	8 ♠	8 ♥	8 ♦
9 ♣	9 ♠	9 ♥	9 ♦
10 ♣	10 ♠	10 ♥	10 ♦
Jack ♣	Jack ♠	Jack♥	Jack ♦
Queen ♣	Queen ♠	Queen ♥	Queen ♦
King ♣	King ♠	King ♥	King ♦

Appendix A, Part 4 - Results for the total of two 6-sided dice

+	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

+	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Appendix B – Glossary and Index

95% confidence statement – Page 120, Section 4.4

A confidence statement is a summary statement of the findings of a study. All confidence statements have the form 'We are 95% confident that the true proportion of (parameter of interest) will be between (low value of confidence interval) and (high value of confidence interval).'

Back to Back Stem Plots – Page 184, Section 5.6

A stem plot in which two sets of numerical data share the stems in the middle, with one set having its leaves going to the right and the other set having its leaves going to the left.

Bar Graph – Page 137, Section 5.1

A graph in which each bar shows how frequently a given category occurs. The bars can go either horizontally or vertically. Bars should be of consistent width and need to be equally spaced apart. The categories may be placed in any order along the axis.

Bias - Page 95, 103, Section 4.1, 4.2

Bias occurs when a measurement repeatedly reports values that are either too high or too low.

Bin Width

See Class Size

Bivariate Data - Page 200, Section 6.1

Numerical data that measures two variables.

Blind Study - Page 126, Section 4.5

A study in which the subject does not know exactly what treatment they are getting.

Block Design - Page 128, Section 4.5

A study in which subjects are divided into distinct categories with certain characteristics (for example, males and females) before being randomly assigned treatments in an experiment.

Box Plot (Box and Whisker Plot) - Page 171, Section 5.5

A display in which a numerical data set is divided into quarters. The 'box' marks the middle 50% of the data and the 'whiskers' mark the upper 25% and lower 25% of the data.

Categorical Variable - Page 93, 136, Section 4.1, 5.1

Variables that can be put into categories, like favorite color, type of car you own, your sports jersey number, etc...

Census - Page 97, 101, Section 4.1, 4.2

A special type of study in which data is gathered from every single member of the population.

Center - Page 147, 156, Section 5.2, 5.3

Typically, it is the mean, median, or the mode of a data set. In a normal distribution curve the mean, median, and mode all mark the center. If a data set is skewed or has outliers, it is standard practice to use the median as the center.

Chance Behavior - Page 26, Section 2.1

Events whose outcomes are not predictable in the short term, but have long term predictability.

Class Size (Bin Width) - Page 164, Section 5.4

A consistent width that all bars on a histogram have. A quick estimation of a reasonable class size is to roughly divide the range by a value from about 7 to 10.

Coincidence - Page 215, Section 6.2

A relationship between two variables that simply occurs by chance.

Combination - Page 15, Section 1.4

An arrangement of a set of objects in which the order does not matter.

$$_nC_r = \frac{n!}{r!(n-r)!}$$

Common Response - Page 214, Section 6.2

A situation in which two variables have a strong correlation but are actually responding to an additional lurking variable.

Complement of an Event - Page 26, Section 2.1

The probability of an event, 'A', NOT occurring. It can be thought of the opposite of an event and can be notated as Ac or A'. P(A') = 1 – P(A)

Compound Event - Page 33, Section 2.2

An event with two or more steps such as drawing a card and then rolling a die.

Conditional Probability - Page 54, Section 2.5

The probability of a particular outcome happening assuming a certain prerequisite condition has already been met. A clue that a conditional probability is being considered is the word 'given' or the vertical bar symbol, |.

Confidence Interval - Page 119, Section 4.4

The range of answers included within the margin of error. Typically, we use a 95% confidence interval meaning it is very likely (95% chance) that the parameter lies within this range.

Confounding - Page 215, Section 6.2

Occurs when two variables are related, but it is not a clear cause/effect relationship because there may be other variables that are influencing the observed effect.

Context - Page 156, 204 Section 5.3

The specific realities of the situation we are considering. We often consider the labels and units when defining the context.

Contingency Table

See Two-Way Table

Control - Page 125, Section 4.5, 6.1

A researcher in an experiment establishes control when one of the treatment groups receives either a placebo or the currently accepted treatment.

Control Group - Page 125, Section 4.5

A group in an experiment that does not receive the actual treatment, but rather receives a placebo or a known treatment.

Convenience Sample - Page 106, Section 4.2

A biased sampling method in which data is only gathered from those individuals who are easy to access or are conveniently located.

Correlation (r) - Page 210-213, Section 6.2

A statistic that is used to measure the strength and direction of a linear correlation whose values range from -1 to 1. The sign of the correlation (+/-) matches the sign of the slope of the regression equation. A correlation value of 0 indicates no linear relationship whatsoever.

Data - Page 93, Section 4.1

A collection of facts, measurements, or observations about a set of individuals.

Density Curve - Page 236, Section 7.1

A curve that gives a rough description of a distribution. The curve is smooth and always has an area equal to 1 or 100%.

Direct Cause and Effect - Page 214, Section 6.2

A situation in which one variable causes a specific effect to occur with no lurking variables.

Direction - Page 210, Section 6.2

One of three general results reported for a linear regression. It will be reported as either being positive, negative, or 0.

Disjoint

See Mutually Exclusive Events

Dot Plot - Page 154, Section 5.3

A simple display that places a dot above each marked value on the x-axis. There is a dot for each result, so results that occur more than once will be shown by stacked dots.

Double Blind - Page 126, Section 4.5

A study in which neither the person administering the treatments nor the subject knows which treatment is being given.

Empirical Rule (68-95-99.7 Rule) - Page 238, Section 7.1

A rule stating that in a normal distribution, 68% of the data is located within one standard deviation of the mean, 95% of the data is located within two standard deviations of the mean, and 99.7% of the data is located within three standard deviations of the mean.

Event - Page 1, Section 1.1

Any action from which a result will be recorded or measured.

Expected Value - Page 67, Section 3.1

The average result over the long run for an event if repeated a large number of times.

Experiment - Page 97, 124, Section 4.1, 4.5

A study in which the researchers impose a treatment on the subjects.

Explanatory Variable - Page 125, 200, Section 4.5, 6.1

The x-axis variable. It can often be viewed as the 'cause' variable or the independent variable.

Factorial - Page 7, Section 1.2

A number followed by an exclamation point indicated repeated multiplication down to 1. For example, $4! = 4 \times 3 \times 2 \times 1$.

Fair Game - Page 76, Section 3.2

A game in which neither the player nor the house has an advantage. An average player over the long run will neither gain nor lose money. In other words, the expected value of the game is the same as the cost to play the game.

Five-Number Summary - Page 171, Section 5.5

A description of data that includes the minimum, first quartile, median, third quartile, and maximum numbers which can be used to create a box plot.

Form - Page 204, Section 6.1

A general description of the pattern in a scatterplot. Typical descriptions include linear, curved, or random (no specific form).

Frequency Table - Page 137, Section 5.1

A table that shows the number of occurrences in each category.

Fundamental Counting Principle - Page 4, Section 1.2

A rule that states that in order to find the number of outcomes for a multi-step event, simply multiply the number of possibilities from each step of the event.

Histogram - Page 164, Section 5.4

A special bar graph for a numerical data set. In a histogram, each bar has the same bin width and there is no space between consecutive bars. Each bar tracks the number or frequency of results in its given range.

Independent Events - Page 33, Section 2.2

Two events are independent if the outcome of one event does not change the probability for the outcome for the other event.

Individual - Page 93, Section 4.1

This is the person, animal, or object being studied.

Interquartile Range (IQR) - Page 174, Section 5.5

The distance between the lower and upper quartiles. $IQR = Q_3 - Q_1$

Instrument of Measurement - Page 94, Section 4.1

This is the tool used to make measurements. Some examples of instruments include rulers, scales, thermometers, or speedometers.

Intersection of Events - Page 42, Section 2.3

In a Venn Diagram, it includes the results that are members of more than one group simultaneously. We use the symbol, ∩, to indicate the intersection and think of the intersection as those parts of the diagram that include both A and B.

Law of Large Numbers - Page 26, 84, Section 2.1, 3.3

A rule that states that we will eventually get closer to the theoretical probability as we greatly increase the number of times an event is repeated.

Line Graph

See Time Plot

Lurking Variable - Page 124, 214, Section 4.5, 6.2

An additional variable that was not taken into account in a particular situation.

Margin of Error - Page 119, Section 4.4

It is the distance we move above and below the mean to help establish a 95% confidence interval in which we believe the true parameter is located. An approximation for the margin of error for a 95% confidence interval is M.O.E $= \pm \frac{1}{\sqrt{n}}$ where n represents the sample size.

Mean (Average) - Page 147, 237, Section 5.2, 7.1

The sum of all the numbers divided by the number of values in a data set. It is also located at the center of a normal distribution and is a good measure of center for symmetric data sets.

Median - Page 147, Section 5.2

The data result in the middle of a data list that has been organized from smallest to largest. If there are two middle data values, then the median is located halfway between those two values. Visually, it marks the spot where half of the area of a graph is below the median and half of the area is above the median. It is common to use the median as your measure of center for skewed data sets or data sets that contain outliers.

Mode - Page 147, Section 5.2

The result that appears most frequently in a data set. It also occurs at the highest point of a density curve.

Multistage Random Sample - Page 104, Section 4.2

A sampling technique that uses randomly selected sub-groups of a population before random selection of individuals occurs.

Mutually Exclusive Events (Disjoint) - Page 41, Section 2.3

Outcomes that cannot occur at the same time. For example, if a single card is drawn from a standard deck, the outcomes of a diamond and a black card are mutually exclusive.

Negative Linear Association - Page 205, Section 6.1

A situation such that as one numerical variable increases, another numerical variable decreases.

Non-Response - Page 108, Section 4.2

A non-sampling error in which individuals selected for a study do not participate or do not answer questions in a survey.

Normal Distribution Curve - Page 237, Section 7.1

A bell-shaped curve that describes a symmetrical data set such that the most frequent results occur near the mean and results become less frequent as you move further from the mean.

Numerical Variable - Page 93, Section 4.1

A variable that can be assigned a numerical value, such as height, distance, or temperature.

Observational Study - Page 97, 124, Section 4.1, 4.5

A study in which researchers do not impose a treatment on the individuals being studied. Data is collected by observing the individuals, surveying the individuals, or collecting data from the individuals from information that is already available. (Observe but do not disturb)

Outcome - Page 1, Section 1.1

A possible result of an event.

Outlier - Page 155, 178, 204, Section 5.3, 5.5, 6.1

A value that is unusual when compared to the rest of a data set. High outliers will be greater than $Q_3 + 1.5$ IQR. Low outliers will be below $Q_1 - 1.5$ IQR.

Parallel Box Plots - Page 183, Section 5.6

Multiple box plots graphed on the same axes to compare multiple data sets.

Parameter - Page 111, Section 4.2

A value that describes the truth about a population. The value is frequently unknown so a parameter is often given as a description of truth.

Permutation - Page 10, Section 1.3

A specific order or arrangement of a set of objects or items. In a permutation, the order in which the items are selected matters.

Pictograph - Page 141, Section 5.1

A bar graph that uses pictures instead of bars. These graphs can be misleading because pictures measure height and width, where bar graphs measure only height. To be effective, all the pictures used must be the same size.

Pie chart - Page 139, Section 5.1

A graph which shows each category as a part of the whole in a circle graph. Pie charts can be used if exactly 100% of the results from a particular situation are known.

Placebo - Page 126, Section 4.5

A fake treatment that is similar in appearance to the real treatment.

Placebo Effect - Page 126, Section 4.5

The placebo effect occurs when a subject starts to experience changes simply because they believe they are receiving a treatment.

Population - Page 101, Section 4.2

The entire group of individuals we are interested in. A population is often described using the word 'all'.

Positive Linear Association - Page 205, Section 6.1

A situation in which as one numerical variable increases, the other numerical variable also increases.

Prime Number - Page 42, Section 2.3

A number that has exactly 2 factors. Remember, 1 is not a prime number!

Probability - Page 26, Section 2.1

The likelihood of a particular outcome occurring.

Probability Model - Page 49, Section 2.4

A table that lists all the values for the outcomes of an event and their respective probabilities. The sum of all the probabilities in a probability model must equal 1.

Processing Errors - Page 109, Section 4.2

An error commonly made due to issues like poor calculations or inaccurate recording of results.

Prospective Studies - Page 124, Section 4.5

A study which follows up with study subjects in the future in an effort to see if there were any long-term effects.

Quartile 1 - Page 172, Section 5.5

The median of all the values to the left of the median. Do not include the median itself in this calculation if the median is one of the data points.

Quartile 3 - Page 172, Section 5.5

The median of all the values to the right of the median. Do not include the median itself in this calculation if the median is one of the data points.

Random Digit Table - Pages 82, 114, Section 3.3, 4.3, Appendix A

A long list of randomly chosen digits from 0 to 9, usually generated by computer software or calculators. A table of random digits can be found in Appendix A, Part 1.

Random Event - Page 26, Section 2.1

An event is random if it does not have short-term predictability but it has long-term predictability. For example, a coin flip is a random event because we do not know what will happen on the next flip, but we can be reasonably sure that about 50% of a long series of flips will land on heads.

Random Sampling Error - Page 107, Section 4.2

Even though a sample is randomly selected, it is entirely possible that a particular result within the population will be over-represented causing us to be significantly different from the parameter. Larger sample sizes reduce random sampling error. The margin of error is stated with most studies to account for random sampling error.

Range - Page 148, 174, Section 5.2, 5.5

A basic description of how spread out a data set is. It is calculated by subtracting the smallest number from the largest number in a data set.

Reliability - Page 95, Section 4.1

How consistently a particular measurement technique gives the same, or nearly the same measurement.

Response Bias - Page 109, Section 4.2

Occurs when an individual responds to a survey with an incorrect or untruthful answer. This type of bias can frequently happen when questions are potentially sensitive or embarrassing.

Response Variable - Page 125, 200, Section 4.5, 6.1

This is the y-axis variable. It can often be thought of as the 'effect' variable or dependent variable.

Retrospective Study - Page 124, Section 4.5

A study in which information about a subject's past is used in the study.

Sample - Page 102, Section 4.2

A representative subset of a population.

Sample Space - Page 1, Section 1.1

A list of all the possible outcomes that may occur.

Sample Survey - Page 97, Section 4.1

A survey that uses a subset of the population in order to try to make predictions about the entire population.

Sampling Frame - Page 103, Section 4.2

A list of all members of a population.

Scatterplot - Page 200, Section 6.1

Graphs that represent a relationship between two numerical variables where each data point is shown as a coordinate point on a scaled grid.

SCOFD - Page 203-206, Section 6.1

This is an acronym used for the description of a scatterplot and stands for Strength, Context, Outliers, Form, and Direction.

Simple Random Sample (SRS) - Page 103, Section 4.2

A sample where all possible groups of a particular size are equally possible. It can be thought of as putting names of all members of a population in a hat and randomly drawing until the desired sample size is reached.

Simulation - Page 82, Section 3.3

A model of a real situation that can be used to make predictions about what might really happen. Often, tables of random digits are used to carry out simulations.

Skewed Distribution - Page 155, 236, Section 5.3, 7.1

A distribution in which the majority of the data is concentrated on one end of the distribution. Visually, there is a 'tail' on the side with less data and this is the direction of the skew.

SOCCS - Page 154-156, Section 5.3

An acronym used to remember the key information to discuss for a distribution: Shape, Outliers, Center, Context, and Spread.

Spread - Page 156, Section 5.3

A way to measure variability of a data set. Common measures of spread are the range, standard deviation, and IQR.

Standard Deviation - Page 174, 237, Section 5.5, 7.1

A measure of spread relative to the mean of a data set. Use this measurement for any data set which is approximately normally distributed.

Statistic - Page 111, Section 4.2

A number that describes results from sample. This number is often a percentage and is used to make an approximation of the parameter.

Stem Plot - Page 157, Section 5.3

A method of organizing data that sorts the data in a visual fashion. The stem is made up of all the leading digits of a piece of data and the leaf is the final digit. No commas or decimal points should be used in a stem plot.

Stratified Random Sample - Page 104, Section 4.2

A sample in which the population is divided into distinct groups called strata before a random sample is chosen from each strata.

Strength - Page 203, 210, Section 6.1, 6.2

One of three measurements reported for a best-fit line that describes how close the data is to being perfectly linear.

Subjects - Page 125, Section 4.5

The individuals that are being studied in an experiment.

Symmetrical Distribution - Page 155, Section 5.3

A distribution in which the left side of the distribution looks like a mirror image of the right side of the distribution.

Systematic Random Sample - Page 104, Section 4.2

A sampling method in which the first selection is made randomly and then a 'system' is used to make the remaining selections. For example, randomly select one person from a list and then select every 14th person after that.

Theoretical Model - Page 26, 82 Section 2.1, 3.3

A model that gives a picture of exactly the frequencies of what should happen in a situation involving probability.

Theoretical Probability - Page 26, Section 2.1

A mathematical calculation of the likelihood that a given outcome will occur.

Time Plot (Line Graph) - Page 145, Section 5.2

A graph that shows how a numerical variable changes over time.

Tree Diagram - Page 2, 4, 48 Section 1.1, 1.2, 2.4

A visual representation of a multi-step event where each successive step branches off from the previous step.

Two-Way Table (Contingency Table) - Page 55, Section 2.5

A table which tracks two characteristics from a set of individuals. For example, we might track gender and grade of all the students in your high school.

Undercoverage - Page 107, Section 4.2

A sampling error in which an entire group or groups of subjects are left out or underrepresented in a study.

Union of Events - Page 41, Section 2.3

A union includes all results that are in either one category, another category, or both categories in a Venn diagram. We use the symbol $\cup$ and can think of a union as anything belonging to either A, B, or both A and B.

Validity - Page 95, Section 4.1

A measurement technique is valid if it is a reasonable way to collect data.

Variables - Page 93, Section 4.1

Characteristics about the individuals in a study in which researchers might have interest.

Venn Diagrams - Page 29, 42, Section 2.1, 2.3

Diagrams that represent outcomes or categories using intersecting circles.

Voluntary Response Survey - Page 105, Section 4.2

A biased sampling method in which participants get to choose whether or not to participate in the survey. The bias occurs because those who are most passionate about an issue will be more likely to respond.

Wording of a Question - Page 108, Section 4.2

The wording of a question can be used to manipulate individuals in a survey such that they are more likely to respond a certain way in the survey which causes bias.

Z-Score - Page 245, Section 7.2

A measure of the number of standard deviations a particular data point is away from the mean in a normal distribution. If a z-score is positive, the value is larger than the mean and if it is negative, it is less than the mean.

Appendix C – Calculator Help

This appendix is not meant to be a full guide for calculators common to students who take this course. Rather, it is intended to highlight some of the locations to access a variety of commands commonly used on a TI-30XS Multiview Scientific Calculator and a TI-84 Plus Graphing Calculator. One online source that can be helpful for those of you with graphing calculator issues can be found on the Prentice Hall website at http://www.prenhall.com/divisions/esm/app/calc_v2/.

Topic 1 - Combinations, Permutations, and Factorials

TI-30 XS Multiview

Access located in the **prb** menu. Enter the value for n, select nCr or nPr, and then enter the value for r.

TI-84 Plus

Access located in the **Math**, **PRB** menu. Enter the value for n, select nCr or nPr, and then enter the value for r

Topic 2 – Random Number Generators

TI-30 XS Multiview

Access located in the **prb** menu. Select rand, enter lowest value, enter highest value.

TI-84 Plus

Access located in the **Math**, **PRB** menu. Select RandInt, enter lowest value, enter highest value, enter number of random values desired.

Topic 3 – Means and Standard Deviations

TI-30 XS Multiview

Enter data into L_1 in the **data** menu. Press **2nd data (stat)** and select 1-Var Stats. Arrow down to find the mean, $\bar{x}$, and the standard deviation, s_x.

TI-84 Plus

Enter data in L1 by selecting **STAT** and **EDIT**. Press STAT and CALC and then select 1-Var Stats. Arrow down to find the mean, $\bar{x}$, and the standard deviation, Sx.

Topic 4 – Correlations, Slopes, and Y-Intercepts

TI-30 XS Multiview

Enter data into L_1 and L_2 in the **data** menu. Press **2nd data (stat)** and select 2-Var Stats for L_1 and L_2. Arrow down to find the slope (a), the y-intercept (b) and the correlation coefficient (r).

TI-84 Plus

Enter data in L1 and L2 by selecting **STAT** and **EDIT**. Press **STAT** and **CALC** and then select LinReg(ax+b). Be sure the Xlist and Ylist are L_1 and L_2. If you wish to store you equation into the Y= menu, press VARS, Y-VARS, Function, and Y_1. If the correlation (r) does not show up, go to **2nd CATALOG** and select DiagnosticOn.

Topic 5 – Normal Distributions

TI-30 XS Multiview

This calculator cannot perform normal distribution calculations.

TI-84 Plus

To find the percent of area in a normal curve, select **2nd DISTR** and select normalcdf(. Enter the lower bound, upper bound, mean, and standard deviation. To find a value from a percentile in a normal distribution, select 2nd DISTR and select invNorm(. Enter the %tile, mean, and standard deviation.

Image References

Random Digit Table http://uwsp.edu/math

Normal Distribution Table http://www.regentsprep.org

TI-30XS Multiview Calculator http://education.ti.com

TI-84 Plus Graphing Calculator http://education.ti.com

Appendix D – Selected Answers

Problem Set 6.1

1a) Yes, Exp. = Semesters, Res. = Credits

1b) No

1c) Yes, Exp. = Years, Res. = Salary

1d) Yes, Exp. = Months, Res. = Apps Downloaded

2) +, Moderately Strong, Linear, No Outliers

3a) Exp. = Minutes, Res. = Depth

3b) Scatterplot

3c) +, Strong, Linear, No Outliers
4a) Exp. = Quality, Res. = Price
4b) +, Moderate, Linear, Several Outliers
5a) Exp. = Months, Res. = Hours
5b) Scatterplot
5c) -, Moderately Strong, Linear, 3 Possible Outliers
6a) Exp. = Elevation, Res. = Temperature
6b) -, Strong, Linear, No Outliers
7) 0.0045
8) 0.0625
9a) 0.0470
9b) 0.4280

Problem Set 6.2

1) Strength, Direction
2) +0.8972
3a) Scatterplot, Exp. = Bed Time, Res. = Wake Time
3b) C
3c) B
4a) No, Common Response
4b) No, Coincidence
4c) No, Common Response
4d) No, Confounding
5) Answers will vary
6) No, Answers will vary
7a) Exp. = Temp, Res. = Visitors
7b) Answers will vary, 0.9
7c) +, Strong, Linear, 1 Outlier

8) 1=E, 2=C, 3=B, 4=A, 5=D
9) Scatterplot, Answers will vary
10) 0.3656
11) 5 or 6
12) 0.03797
13) 0.00077

Problem Set 6.3

1a) Scatterplot
1b) $\hat{y}$ = 2.3818 + 1.9795x
1c) +0.9950
1d) Slope = 1.9795
1e) 36 cm, 121 cm
2a) Scatterplot
2b) +, Strong, Curved, No Outliers
2c) $\hat{y}$ = -181.5867 + 0.0933x
2d) +0.9624
2e) $6.51, Not Accurate
2f) $2.03, Too High by $0.43
3a) Exp. = Father's IQ, Res. = Son's IQ
3b) Slope = 0.9
3c) Y-Intercept = 12
3d) Slope is Reasonable, Y-Intercept is Not
3e) 120, 138
3f) Answers will vary
4a) Scatterplot, $\hat{y}$ = 7.757 – 0.478x
4b) Slope = -0.478
4c) Y-Intercept = 7.757
4d) -0.8564
4e) 2.021 hours
4f) -2.6 months
5a) Exp. = Absences, Res. = Grade
5b) -, Strong, Linear, No Outliers
5c) 50.354
5d) 20.582
5e) -0.9345
5f) Answers will vary
6a) $\hat{y}$ = 2.419 + 1.884x, r = +0.9858
6b) $\hat{y}$ = 10.2 - 0.2x, r = -0.1129
6c) Answers will vary
7a) – 7f) Answers will vary
8a) $\hat{y}$ = -2714.859+35.078x
8b) 35.078
8c) 0.7823
8d) 477, -1136, Answers vary
9) 0.4930
10) 0.3932
11) 0.6105
12) C
13a) mean = $10.76, Standard Deviation = $2.69
13b) $9.25, $9.45, $9.80, $10.60, $18.90
13c) Box Plot
13d) Median, 5# Summary
13e) Skewed Right, 1 Outlier, Median = $7.80, IQR = $1.15

Chapter 6 Review

1) False
2) True
3) True
4) False
5) r
6) Scatterplot
7) Explanatory
8) -1 to 1
9) The Slope
10) Least Squares Regression Line
11) -1 or 1
12) 0.7396
13) +0.8775 or -0.8775
14) Extrapolation
15) Interpolation
16) Strength, Context, Outliers, Form, and Direction
17) Common Response
18a) No. Coincidence
18b) No. Answers will vary
18c) No. Answers will vary
19a) Scatterplot, Exp. = Beers, Res. = BAC
19b) $\hat{y}$ = -0.0215+0.0259x, r=+0.8209
19c) Slope = 0.0259
19d) Y-Intercept = -0.0215
19e) 0.1339
19f) 0.367
19g) Answers will vary
19h) 5 to 6
20a) Scatterplot, Exp. = Skid Length, Res. = Speed
20b) $\hat{y}$ = 18.825 + 0.2341x
20c) +, Strong, Curved, No Outliers
20d) +0.9805, Answers vary
20e) 55.6 mph
20f) 27.3 mph
20g) Overestimate