Chapter 2: Statistics: Part 2

Graphical descriptions of data are important. However, many times we want to have a number to help describe a data set. As an example, in baseball a pitcher is considered good if he has a low number of earned runs per nine innings. A baseball hitter is considered good if he has a high batting average. These numbers tell us a great deal about a player. There are similar numbers in other sports such as percentage of field goals made in football. There are also similar numbers in other aspects of life. If you want to know how much money you will make when you graduate from college and are employed in your chosen field, you could look at the average salary that someone with your degree earns. If you want to know if you can afford to purchase a home, you could look at the median price of homes in the area. To understand how to find this information, we need to look at the different numerical descriptive statistics that exist out there.

Numerical Descriptive Statistics: These are numbers that are calculated from the sample and are used to describe or estimate the population parameter.

Statistics that we can calculate are proportion, location of center (average), measures of spread (variability), and percentiles. There are other numbers, but these are the ones that we will concentrate on in this book.

Section 2.1: Proportion

Proportions are usually calculated when dealing with qualitative variables. Suppose that you want to know the proportion of time that a basketball player will make a free throw. You could look at how often the player tries to make the free throw, and how often they do make a free throw. Then you could divide the number made by the number attempted. This is how we find proportion. This is a sample statistic, since we cannot look at all of the attempts, because the player could attempt more in the future. If the player retires, and never wants to play basketball ever again, then we could find the population parameter for that player. Since there are rare cases where you can find this, then we will define both the population parameter and the sample statistic. Remember though, usually we use the sample statistic to estimate the population parameter.

Population Proportion:

$$p = \frac{r}{N}$$

where r = number of successes observed
N = number of times the activity could be tried

Sample Proportion:

$$\hat{p} = \frac{r}{n}$$

where r = number of successes observed
n = number of times the activity was tried

Example 2.1.1: Finding Proportion

Suppose that you ask 140 people if they prefer vanilla ice cream to other flavors, and 86 say yes. What is the proportion of people who prefer vanilla ice cream?

Since you only asked 140 people, and there are many more than 140 people in the world, then this is a sample and we use the sample proportion formula.

$$\hat{p} = \frac{r}{n}$$

$r = 86$

$n = 140$

$$\hat{p} = \frac{86}{140} \approx 0.614 = 61.4\%$$

So 61.4% of the people in the sample like vanilla ice cream. This could mean that 61.4% of all people in the world like vanilla ice cream. We do not know for sure, but this is a good guess for the true proportion, p, as long as our sample was representative of the population. If you own an ice cream shop, then you probably want to make sure you order more vanilla ice cream than other flavors.

Section 2.2: Location of Center

The center of a population is very important. This describes where you expect to find values. If you know that you expect to make $50,000 annually when you graduate from college and are employed in your field of study, then that is the location of the center. It does not mean everyone will make that amount. It just means that you will make around that amount. The location of center is also known as the average. There are three types of averages—mean, median, and mode.

Mean: The mean is the type of average that most people commonly call "the average." You take all of the data values, find their sum, and then divide by the number of data values. Again, you will be using the sample statistic to estimate the population parameter, so we need formulas and symbols for each of these.

Population Mean:

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum x}{N}$$

where N = size of the population

$x_1, x_2, \cdots, x_N$ are data values

Note: $\sum x = x_1 + x_2 + \cdots + x_N$ *is a short cut way to write adding a bunch of numbers together*

Sample Mean:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n}$$

where n = size of the sample

$x_1, x_2, \cdots, x_n$ are data values

Note: $\sum x = x_1 + x_2 + \cdots + x_n$ *is a short cut way to write adding a bunch of numbers together*

Median: This is the value that is found in the middle of the ordered data set.

Most books give a long explanation of how to find the median. The easiest thing to do is to put the numbers in order and then count from both sides in, one data value at a time, until you get to the middle. If there is one middle data value, then that is the median. If there are two middle data values, then the median is the mean of those two data values. If you have a really large data set, then you will be using technology to find the value. There is no symbol or formula for median, neither population nor sample.

Mode: This is the data value that occurs most often.

The mode is the only average that can be found on qualitative variables, since you are just looking for the data value with the highest frequency. The mode is not used very often otherwise. There is no symbol or formula for mode, neither population nor sample. Unlike the other two averages, there can be more than one mode or there could be no mode. If you have two modes, it is called bimodal. If there are three modes, then it is called trimodal. If you have more than three modes, then there is no mode. You can also have a data set where no values occur most often, in which case there is no mode.

Example 2.2.1: Finding the Mean, Median, and Mode (Odd Number of Data Values)

The first 11 days of May 2013 in Flagstaff, AZ, had the following high temperatures (in °F)

Table 2.2.1: Weather Data for Flagstaff, AZ, in May 2013

71	59	69	68	63	57
57	57	57	65	67	

(Weather Underground, n.d.)

Find the mean, median, and mode for the high temperature
Since there are only 11 days, then this is a sample.

Mean:

$$\bar{x} = \frac{71+59+69+68+63+57+57+57+57+65+67}{11}$$

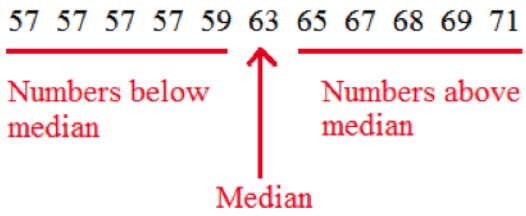
$$= \frac{690}{11}$$

$$\approx 62.7^{\circ}F$$

Median:

First put the data in order from smallest to largest.
57, 57, 57, 57, 59, 63, 65, 67, 68, 69, 71
Now work from the outside in, until you get to the middle number.

Figure 2.2.2: Finding the Median.



So the median is 63°F

Mode:

From the ordered list it is easy to see that 57 occurs four times and no other data values occur that often. So the mode is 57°F.

We can now say that the expected high temperature in early May in Flagstaff, Arizona is around 63°F.

Example 2.2.2: Finding the Mean, Median, and Mode (Even Number of Data Values)

Now let's look at the first 12 days of May 2013 in Flagstaff, AZ. The following is the high temperatures (in °F)

Table 2.2.3: Weather Data for Flagstaff, AZ, in May 2013

71	59	69	68	63	57
57	57	57	65	67	73

(Weather Underground, n.d.)

Find the mean, median, and mode for the high temperature
Since there are only 12 days, then this is a sample.

Mean:

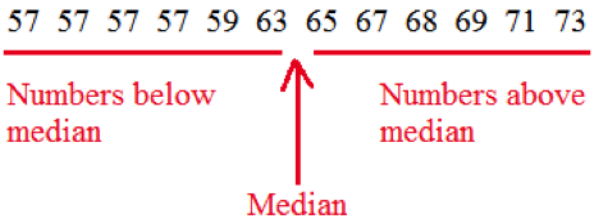
$$\bar{x} = \frac{71+59+69+68+63+57+57+57+57+65+67+73}{12}$$

$$= \frac{763}{12}$$

$$\approx 63.6°F$$

Median:
First put the data in order from smallest to largest.
57, 57, 57, 57, 59, 63, 65, 67, 68, 69, 71, 73
Now work from the outside in, until you get to the middle number.

Figure 2.2.4: Finding the Median



This time there are two numbers that are in the middle. So the median is

$$\text{median} = \frac{63+65}{2} = 64°F.$$

Mode:
From the ordered list it is easy to see that 57 occurs 4 times and no other data values occurs that often. So the mode is 57°F.

Example 2.2.3: Effect of Extreme Values on the Mean and Median

A random sample of unemployment rates for 10 countries in the European Union (EU) for March 2013 is given:

Table 2.2.5: Unemployment Rates for EU Countries

11.0	7.2	13.1	26.7	5.7	9.9	11.5	8.1	4.7	14.5

(Eurostat, n.d.)

Find the mean, median, and mode.
Since the problem says that it is a random sample, we know this is a sample. Also, there are more than 10 countries in the EU.

Mean:

$$\bar{x} = \frac{11.0 + 7.2 + 13.1 + 26.7 + \cdots + 14.5}{10}$$
$$= \frac{112.4}{10}$$
$$= 11.24$$

The mean is 11.24%.

Median:

4.7, 5.7, 7.2, 8.1, 9.9, 11.0, 11.5, 13.1, 14.5, 26.7
Both 9.9 and 11.0 are the middle numbers, so the median is

$$\text{median} = \frac{9.9 + 11.0}{2} = 10.45$$

The median is 10.45%.

Note: This data set has no mode since there is no number that occurs most often.

Now suppose that you remove the 26.7 from your sample since it is such a large number (an outlier). Find the mean, median, and mode.

Table 2.2.6: Unemployment Rates for EU Countries

11.0	7.2	13.1	5.7	9.9	11.5	8.1	4.7	14.5

$$\bar{x} = \frac{11.0 + 7.2 + 13.1 + 5.7 + \cdots + 14.5}{9}$$
$$= \frac{85.7}{9}$$
$$\approx 9.52$$

The mean is 9.52%
The median is 9.9%.
There is still no mode.

Notice that the mean and median with the 26.7 were a bit different from each other. When the 26.7 value was removed, the mean dropped significantly, while the median dropped, but not as much. This is because the mean is affected by extreme values called outliers, but the median is not affected by outliers as much.

In section 1.5, there was a discussion on histogram shapes. If you look back at Graphs 1.5.11, 1.5.12, and 1.5.13, you will see examples of symmetric, skewed right, and skewed left graphs. Since symmetric graphs have their extremes equally on both sides, then the mean would not be pulled in any direction, so the mean and the median are essentially the same value. With a skewed right graph, there are extreme values on the right, and they will pull the mean up, but not affect the median much. So the mean will be higher than the median in skewed right graphs. Skewed left graphs have their extremes on the left, so the mean will be lower than the median in skewed left graphs.

Example 2.2.4: Finding the Average of a Qualitative Variable

Suppose a class was asked what their favorite soft drink is and the following is the results:

Table 2.2.7: Favorite Soft Drink

Coke	Pepsi	Mt. Dew	Coke	Pepsi	Dr. Pepper	Sprite	Coke	Mt. Dew
Pepsi	Pepsi	Dr. Pepper	Coke	Sprite	Mt. Dew	Pepsi	Dr. Pepper	Coke
Pepsi	Mt. Dew	Coke	Pepsi	Pepsi	Dr. Pepper	Sprite	Pepsi	Coke
Dr. Pepper	Mt. Dew	Sprite	Coke	Coke	Pepsi			

Find the average.

Remember, mean, median, and mode are all examples of averages. However since the data is qualitative, you cannot find the mean and the median. The only average you can find is the mode. Notice, Coke was preferred by 9 people, Pepsi was preferred by 10 people, Mt Dew was preferred by 5 people, Dr. Pepper was preferred by 5 people, and Sprite was preferred by 4 people. So Pepsi has the

highest frequency, so Pepsi is the mode. If one more person came in the room and said that they preferred Coke, then Pepsi and Coke would both have a frequency of 10. So both Pepsi and Coke would be the modes, and we would call this bimodal.

Section 2.3: Measures of Spread

The location of the center of a data set is important, but also important is how much variability or spread there is in the data. If a teacher gives an exam and tells you that the mean score was 75% that might make you happy. But then if the teacher says that the spread was only 2%, then that means that most people had grades around 75%. So most likely you have a C on the exam. If instead you are told that the spread was 15%, then there is a chance that you have an A on the exam. Of course, there is also a chance that you have an F on the exam. So the higher spread may be good and it may be bad. However, without that information you only have part of the picture of the exam scores. So figuring out the spread or variability is useful.

Measures of Spread or Variability: These values describe how spread out a data set is.

There are different ways to calculate a measure of spread. One is called the range and another is called the standard deviation. Let's look at the range first.

Range: To find the range, subtract the minimum data value from the maximum data value. Some people give the range by just listing the minimum data value and the maximum data value. However, to statisticians the range is a single number. So you want to actually calculate the difference.

$$\text{Range} = \text{maximum} - \text{minimum}$$

The range is relatively easy to calculate, which is good. However, because of this simplicity it does not tell the entire story. Two data sets can have the same range, but one can have much more variability in the data while the other has much less.

Example 2.3.1: Finding the Range

Find the range for each data set.

a. 10, 20, 30, 40, 50

$\text{Range} = 50 - 10 = 40$

b. 10, 35, 36, 37, 50

$\text{Range} = 50 - 10 = 40$

Notice both data sets from Example 2.3.1 have the same range. However, the one in part b seems to have most of the data closer together, except for the extremes. There seems to be less variability in the data set in part b than in the data set in part a. So we need a better way to quantify the spread.

Instead of looking at the difference between highest and lowest, let's look at the difference between each data value and the center. The center we will use is the mean. The difference between the data value and the mean is called the deviation.

Deviation from the Mean: $\text{data value} - \text{mean} = x - \bar{x}$

To see how this works, let's use the data set from Example 2.2.1. The mean was about 62.7°F

Table 2.3.1: Finding the Deviations

x	$x - \bar{x}$
71	8.3
59	-3.7
69	6.3
68	5.3
63	0.3
57	-5.7
57	-5.7
57	-5.7
57	-5.7
65	2.3
67	4.3
Sum	**0.3**

Notice that the sum of the deviations is around zero. If there is no rounding of the mean, then this should add up to exactly zero. So what does that mean? Does this imply that on average the data values are zero distance from the mean? No. It just means that some of the data values are above the mean and some are below the mean. The negative deviations are for data values that are below the mean and the positive deviations are for data values that are above the mean. So we need to get rid of the sign (positive or negative). How do we get rid of a negative sign? Squaring a number is a widely accepted way to make all of the numbers positive. So let's square all of the deviations.

Squared Deviations from the Mean: To find these values, square the deviations from the mean. Also, you can think of this as being the squared distance from the mean.

So, for the data set, let's find the squared deviations.

Table 2.3.2: Finding the Squared Deviations.

x	$x-\bar{x}$	$(x-\bar{x})^2$
71	8.3	68.89
59	-3.7	13.69
69	6.3	39.69
68	5.3	28.09
63	0.3	0.09
57	-5.7	32.49
57	-5.7	32.49
57	-5.7	32.49
57	-5.7	32.49
65	2.3	5.29
67	4.3	18.49
Sum	**0.3**	**304.19**

Now that we have the sum of the squared deviations, we should find the mean of these values. However, since this is a sample, the normal way to find the mean, summing and dividing by n, does not estimate the true population value correctly. It would underestimate the true value. So, to calculate a better estimate, we will divide by a slightly smaller number, $n-1$. This strange average is known as the sample variance.

Sample Variance: This is the sum of the squared deviations from the mean divided by $n-1$. The symbol for sample variance is s^2 and the formula for the sample variance is:

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$

For this data set, the sample variance is

$$s^2 = \frac{304.19}{11-1} = \frac{304.19}{10} = 30.419$$

The variance measures the average squared distance from the mean. Since we want to know the average distance from the mean, we will need to take the square root at this point.

Sample Standard Deviation: This is the square root of the variance. The standard deviation is a measure of the average distance the data values are from the mean. The symbol for sample standard deviation is s and the formula for the sample standard deviation is

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

Thus, for this data set, the sample standard deviation is $s = \sqrt{30.419} \approx 5.52^{\circ}F$.
Note: The units are the same as the original data.

Since the sample variance and the sample standard deviation are used to estimate the population variance and population standard deviation, we should define the symbols and formulas for those as well.

Population Variance: $\sigma^2 = \frac{\sum (x - \mu)^2}{N}$

Population Standard Deviation: $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x - \mu)^2}{N}}$

Example 2.3.2: Finding the Range, Variance, and Standard Deviation

A random sample of unemployment rates for 10 counties in the EU for March 2013 is given

Table 2.3.3: Unemployment Rates for EU Countries

11.0	7.2	13.1	26.7	5.7	9.9	11.5	8.1	4.7	14.5

(Eurostat, n.d.)
Find the range, variance, and standard deviation.

Since this is a sample, then we will use the sample statistics formulas.
In Example 2.2.3, we calculated the mean to be 11.24%. The maximum value is 26.7% and the minimum value is 4.7%. So the range is:

range = 26.7 – 4.7 = 22.0%

To find the variance and the standard deviation, it is easier to use a table then the formula. The table follows the formula though, so they are the same thing.

Table 2.3.4: Finding the Variance and Standard Deviation

x	$x-\bar{x}$	$(x-\bar{x})^2$
11.0	-0.24	0.0576
7.2	-4.04	16.3216
13.1	1.86	3.4596
26.7	15.46	239.0116
5.7	-5.54	30.6916
9.9	-1.34	1.7956
11.5	0.26	0.0676
8.1	-3.14	9.8596
4.7	-6.54	42.7716
14.5	3.26	10.6276
Sum	**0**	**354.664**

Sample variance:

$$s^2 = \frac{354.664}{10-1} = \frac{354.664}{9} \approx 39.40711111$$

Sample standard deviation:

$$s = \sqrt{39.4071111} \approx 6.28\%$$

So, the unemployment rates for countries in the EU are approximately 11.24% with an average spread of about 6.28%. Since the sample standard deviation is fairly high compared to the mean, then there is a great deal of variability in unemployment rates for countries in the EU. This means that countries in the EU have rates that are much lower than the mean and some that have rates much higher than the mean.

Percentiles

There are other calculations that we can do to look at spread. One of those is called percentile. This looks at what data value has a certain percent of the data at or below it.

Percentiles: A value with k-percent of the data at or below this value.

For example, if a data value is in the 80th percentile, then 80% of the data values fall at or below this value.

We see percentiles in many places in our lives. If you take any standardized tests, your score is given as a percentile. If you take your child to the doctor, their height and weight are given as percentiles. If your child is tested for gifted or behavior problems, the score

is given as a percentile. If your child has a score on a gifted test that is in the 92nd percentile, then that means that 92% of all of the children who took the same gifted test scored the same or lower than your child. That also means that 8% scored the same or higher than your child. This may mean that your child is gifted.

Example 2.3.3: Interpreting Percentiles

Suppose you took the SAT mathematics test and received your score as a percentile.

a. What does a score in the 90th percentile mean?
   90 percent of the scores were at or below your score (You did the same as or better than 90% of the test takers.)

b. What does a score in the 70th percentile mean?
   70% of the scores were at or below your score.

c. If the test was out of 800 points and you scored in the 80th percentile, what was your score on the test?
   You do not know! All you know is that you scored the same as or better than 80% of the people who took the test. If all the scores were really low, you could have still failed the test. On the other hand, if many of the scores were high you could have gotten a 95% on the test.

d. If your score was in the 95th percentile, does that mean you passed the test?
   No, it just means you did the same as or better than 95% of the other people who took the test. You could have failed the test, but still did the same as or better than 95% of the rest of the people.

There are three percentiles that are commonly used. They are the first, second, and third quartiles, where the quartiles divide the data into 25% sections.

First Quartile (Q_1): 25th percentile (25% of the data falls at or below this value.)
Second Quartile (Q_2 or M): 50th percentile, also known as the median (50% of the data falls at or below this value.).
Third Quartile (Q_3): 75th percentile (75% of the data falls at or below this value.)

To find the quartiles of a data set:
Step 1: Sort the data set from the smallest value to the largest value.
Step 2: Find the median (M or Q_2).
Step 3: Find the median of the lower 50% of the data values. This is the first quartile (Q_1).

Step 4: Find the median of the upper 50% of the data values. This is the third quartile (Q_3).

If we put the three quartiles together with the maximum and minimum values, then we have five numbers that describe the data set. This is called the five-number summary.

Five-Number Summary: Lowest data value known as the minimum (Min), the first quartile (Q_1), the median (M or Q_2), the third quartile (Q_3), and the highest data value known as the maximum (Max).

Also, since we have the quartiles, we can talk about how much spread there is between the 1st and 3rd quartiles. This is known as the interquartile range.

Interquartile Range (IQR): $IQR = Q_3 - Q_1$

There are times when we want to look at the five-number summary in a graphical representation. This is known as a box-and-whiskers plot or a box plot.

Box Plot: Plot of the five-number summary

A box plot is created by first setting a scale (number line) as a guideline for the box plot. Then, draw a rectangle that spans from Q_1 to Q_3 above the number line. Mark the median with a vertical line through the rectangle. Next, draw dots for the minimum and maximum points to the sides of the rectangle. Finally, draw lines from the sides of the rectangle out to the dots.

Example 2.3.4: Find the Five-Number Summary and IQR and Draw a Box Plot (Odd Number of Data Points)

The first 11 days of May 2013 in Flagstaff, AZ, had the following high temperatures (in °F):

Table 2.3.5: Weather Data for Flagstaff, AZ, in May 2013

71	59	69	68	63	57
57	57	57	65	67	

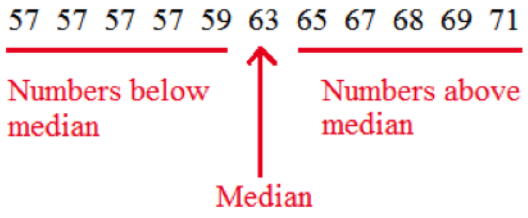
(Weather Underground, n.d.)

Find the five-number summary and IQR and draw a box plot.

To find the five-number summary, you must first put the numbers in order from smallest to largest.
57, 57, 57, 57, 59, 63, 65, 67, 68, 69, 71

Then find the median. The number 63 is in the middle of the data set, so the median is 63°F. To find Q_1, look at the numbers below the median. Since 63 is the median, you do not include that in the listing of the numbers below the median. To find Q_3, look at the numbers above the median. Since 63 is the median, you do not include that in the listing of the numbers above the median.

Figure 2.3.6: Finding the median, Q_1, and Q_3

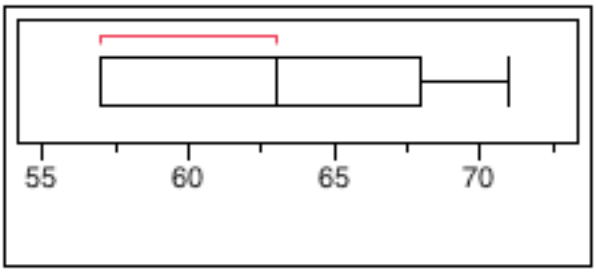


Looking at the numbers below the median, the median of those is 57. $Q_1 = 57$°F. Looking at the numbers above the median, the median of those is 68. $Q_3 = 68$°F. Now find the minimum and maximum. The minimum is 57°F and the maximum is 71°F. Thus, the five-number summary is:
Min = 57°F
$Q_1 = 57$°F
Med = $Q_2 = 63$°F
$Q_3 = 68$°F
Max = 71°F.
Also, the IQR = $Q_3 - Q_1 = 68 - 57 = 11$°F

Finally, draw a box plot for this data set as follows:

Figure 2.3.7: Box Plot

Temperatures in °F in Flagstaff, AZ, in early May 2013



Notice that the median is basically in the center of the box, which implies that the data is not skewed. However, the minimum value is the same as Q_1, so that implies there might be a little skewing, though not much.

Example 2.3.5: Find the Five-Number Summary and IQR and Draw a Box Plot (Even Number of Data Points)

The first 12 days of May 2013 in Flagstaff, AZ, had the following high temperatures (in °F):

Table 2.3.8: Weather Data for Flagstaff, AZ, in May 2013

71	59	69	68	63	57
57	57	57	65	67	73

(Weather Underground, n.d.)

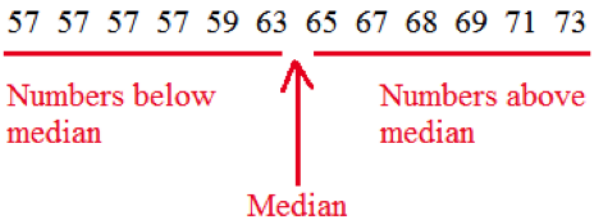
Find the five-number summary and IQR and draw a box plot.
To find the five-number summary, you must first put the data values in order from smallest to largest. 57, 57, 57, 57, 59, 63, 65, 67, 68, 69, 71, 73

Then find the median. The numbers 63 and 65 are in the middle of the data set, so the median is $\frac{63+65}{2} = 64^{\circ} F$

To find Q_1, look at the numbers below the median. Since the number 64 is the median, you include all the numbers below 64, including the 63 that you used to find the median.

To find Q_3, look at the numbers above the median. Since the number 64 is the median, you include all the numbers above 64, including the 65 that you used to find the median.

Figure 2.3.9: Finding the Median, Q_1, and Q_3.



Looking at the numbers below the median (57, 57, 57, 57, 59, 63), the median of those is $\frac{57+57}{2} = 57^{\circ}F$. $Q_1 = 57°F$. Looking at the numbers above the median (65, 67, 68, 69, 71, 73), the median of those is $\frac{68+69}{2} = 68.5^{\circ}F$. $Q_3 = 68.5°F$.

Now find the minimum and maximum. The minimum is 57°F and the maximum is 73°F.

Thus, the five-number summary is:

Min = 57°F

Q_1 = 57°F

Med = Q_2 = 64°F

Q_3 = 68.5°F

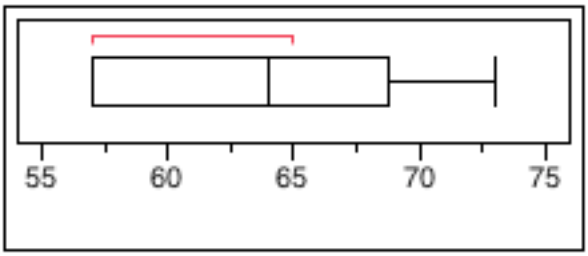
Max = 73°F.

Also, the IQR = $Q_3 - Q_1$ = 68.5 – 57 = 11.5°F

Finally, draw a box plot for this data set as follows:

Figure 2.3.10: Box Plot

Temperatures in °F in Flagstaff, AZ, in early May 2013



Notice that the median is basically in the center of the box, so that implies that the data is not skewed. However, the minimum value is the same as Q_1, so that implies there might be a little skewing, though not much.

It is important to understand how to find all descriptive statistics by hand and also by using a calculator.

Example 2.3.6: Finding the Descriptive Statistics Using the TI-83/84 Calculator

The first 11 days of May 2013 in Flagstaff, AZ, had the following high temperatures (in °F):

Table 2.3.11: Weather Data for Flagstaff, AZ, in May 2013

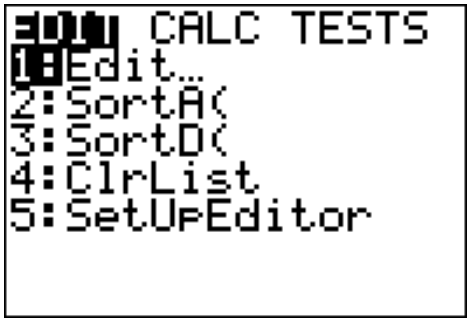
71	59	69	68	63	57
57	57	57	65	67	

(Weather Underground, n.d.)

Find the descriptive statistics for this data set using the TI-83/84 calculator.

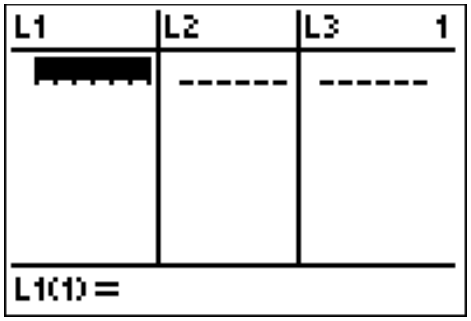
First you need to put the data into the calculator. To do this, press STAT. The STAT button is in the third row of buttons, next to the arrow keys.
Once you press STAT, you will see the following screen:

Figure 2.3.12: STAT Window



Choose 1:Edit… and you will see the following:

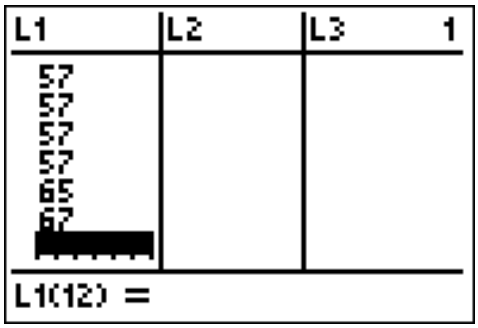
Figure 2.3.13: Edit Window



Note: If there is already data in list 1 (L1), then you should move the cursor up to L1 by using the arrow keys. Then, press clear and enter. This should clear all data from list 1 (L1).

Now type all of the data into list 1 (L1):

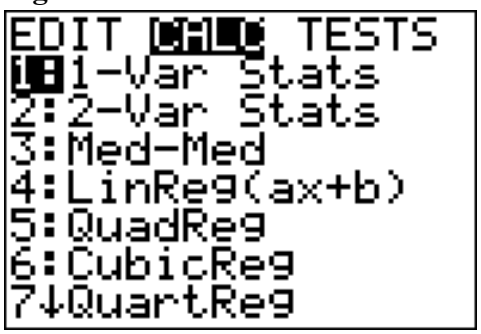
Figure 2.3.14: Data Typed Into L1



Note: Figure 2.3.14 only shows the last six data points entered, but all the data has been entered.

Next, press STAT again and move over to CALC using the right arrow button. You will see the following:

Figure 2.3.15: CALC Window

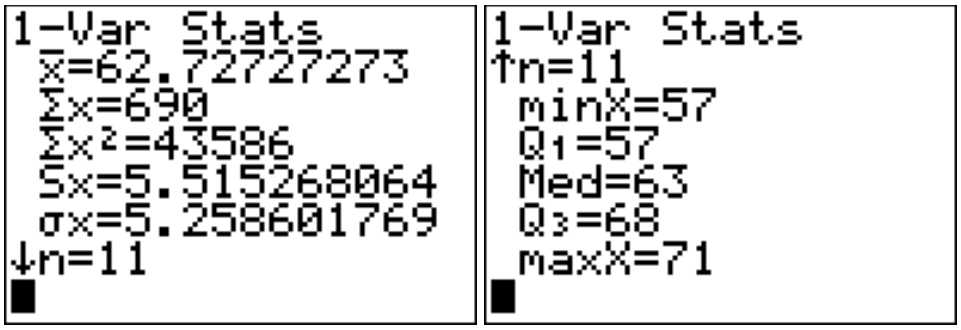


Choose 1:1-Var Stats. This will put 1-Var Stats on your home screen. Type in L1 (2nd 1), and the calculator will show the following:

Figure 2.3.16: 1-Var Stat on Home Screen

At this point press ENTER, and you will see the following: (Use the down arrow button to see the rest of the results.)

Figure 2.3.17: 1-Var Stat Results



Therefore, the mean is $\bar{x} = 62.7°F$, the standard deviation is $s = 5.515°F$, and the five-number summary is Min = 57°F, Q_1 = 57°F, Med = Q_2 = 63°F, Q_3 = 68°F, Max = 71°F. You can find the range by subtracting the max and min. You can find IQR by subtracting Q_3 and Q_1, and you can find the variance by squaring the standard deviation. You cannot find the mode from the calculator. Note that the calculator gives you the population standard deviation $\sigma = 5.259°F$. Notice it is different than the value for s, since they are calculated differently. The value the calculator gives you for the population standard deviation is not the actual true value. The calculator gives you both values because it does not know if you typed in a sample or a population. You can ignore the population standard deviation σ in almost all cases.

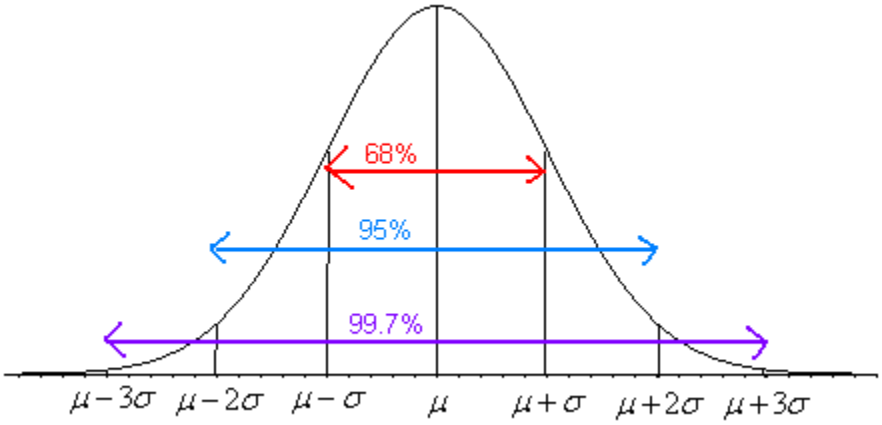
Section 2.4: The Normal Distribution

There are many different types of distributions (shapes) of quantitative data. In section 1.5 we looked at different histograms and described the shapes of them as symmetric, skewed left, and skewed right. There is a special symmetric shaped distribution called the normal distribution. It is high in the middle and then goes down quickly and equally on both ends. It looks like a bell, so sometimes it is called a bell curve. One property of the normal distribution is that it is symmetric about the mean. Another property has to do with what percentage of the data falls within certain standard deviations of the mean. This property is defined as the Empirical Rule.

The Empirical Rule: Given a data set that is approximately normally distributed:
Approximately 68% of the data is within one standard deviation of the mean.
Approximately 95% of the data is within two standard deviations of the mean.
Approximately 99.7% of the data is within three standard deviations of the mean.

To visualize these percentages, see the following figure.

Figure 2.4.1: Empirical Rule



Note: The empirical rule is only true for approximately normal distributions.

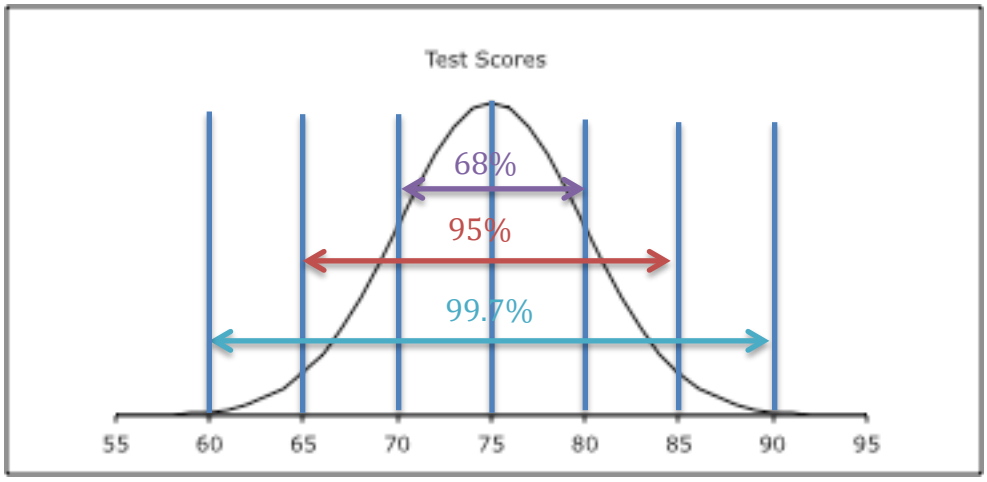
Example 2.4.1: Using the Empirical Rule

Suppose that your class took a test and the mean score was 75% and the standard deviation was 5%. If the test scores follow an approximately normal distribution, answer the following questions:

a. What percentage of the students had scores between 65 and 85?
b. What percentage of the students had scores between 65 and 75?
c. What percentage of the students had scores between 70 and 80?
d. What percentage of the students had scores above 85?

To solve each of these, it would be helpful to draw the normal curve that follows this situation. The mean is 75, so the center is 75. The standard deviation is 5, so for each line above the mean add 5 and for each line below the mean subtract 5. The graph looks like the following:

Figure 2.4.2: Empirical Rule for Example 2.4.1



a. From the graph we can see that 95% of the students had scores between 65 and 85.
b. The scores of 65 to 75 are half of the area of the graph from 65 to 85. Because of symmetry, that means that the percentage for 65 to 85 is ½ of the 95%, which is 47.5%.
c. From the graph we can see that 68% of the students had scores between 70 and 80.
d. For this problem we need a bit of math. If you looked at the entire curve, you would say that 100% of all of the test scores fall under it. So because of symmetry 50% of the test scores fall in the area above the mean and 50% of the test scores fall in the area below the mean. We know from part b that the percentage from 65 to 75 is 47.5%. Because of symmetry, the percentage from 75 to 85 is also 47.5%. So the percentage above 85 is 50% - 47.5% = 2.5%.

When we look at Example 2.4.1, we realize that the numbers on the scale are not as important as how many standard deviations a number is from the mean. As an example, the number 80 is one standard deviation from the mean. The number 65 is 2 standard deviations from the mean. However, 80 is above the mean and 65 is below the mean. Suppose we wanted to know how many standard deviations the number 82 is from the mean. How would we do that? The other numbers were easier because they were a whole number of standard deviations from the mean. We need a way to quantify this. We will use a z-score (also known as a z-value or standardized score) to measure how many standard deviations a data value is from the mean. This is defined as:

z-score: $z = \dfrac{x - \mu}{\sigma}$

where x = data value (raw score)

z = standardized value (z-score or z-value)

μ = population mean

σ = population standard deviation

Note: Remember that the z-score is always how many standard deviations a data value is from the mean of the distribution.

Suppose a data value has a z-score of 2.13. This tells us two things. First, it says that the data value is above the mean, since it is positive. Second, it tells us that you have to add more than two standard deviations to the mean to get to this value. Since most data (95%) is within two standard deviations, then anything outside this range would be considered a strange or unusual value. A z-score of 2.13 is outside this range so it is an unusual value. As another example, suppose a data value has a z-score of -1.34. This data value must be below the mean, since the z-score is negative, and you need to subtract more than one standard deviation from the mean to get to this value. Since this is within two standard deviations, it is an ordinary value.

An **unusual value** has a z-score ≤ -2 or a z-score ≥ 2

A **usual value** has a z-score between -2 and 2, that is $-2 \leq z - score \leq 2$.

You may encounter standardized scores on reports for standardized tests or behavior tests as mentioned previously.

Example 2.4.2: Calculating Z-Scores

Suppose that your class took a test the mean score was 75% and the standard deviation was 5%. If test scores follow an approximately normal distribution, answer the following questions:

a. If a student earned 87 on the test, what is that student's z-score and what does it mean?

$\mu = 75$, $\sigma = 5$, and $x = 87$

$$z = \frac{x - \mu}{\sigma}$$
$$= \frac{87 - 75}{5}$$
$$= 2.40$$

This means that the score of 87 is more than two standard deviations above the mean, and so it is considered to be an unusual score.

b. If a student earned 73 on the test, what is that student's z-score and what does it mean?

$\mu = 75$, $\sigma = 5$, and $x = 73$

$$z = \frac{x - \mu}{\sigma}$$
$$= \frac{73 - 75}{5}$$
$$= -0.40$$

This means that the score of 73 is less than one-half of a standard deviation below the mean. It is considered to be a usual or ordinary score.

c. If a student earned 54 on the test, what is that student's z-score and what does it mean?

$\mu = 75$, $\sigma = 5$, and $x = 54$

$$z = \frac{x - \mu}{\sigma}$$
$$= \frac{54 - 75}{5}$$
$$= -4.20$$

The means that the score of 54 is more than four standard deviations below the mean, and so it is considered to be an unusual score.

d. If a student has a z-score of 1.43, what actual score did she get on the test?

$\mu = 75$, $\sigma = 5$, and $z = 1.43$

This problem involves a little bit of algebra. Do not worry, it is not that hard. Since you are now looking for x instead of z, rearrange the equation solving for x as follows:

$$z = \frac{x - \mu}{\sigma}$$
$$z\cancel{\sigma} = \frac{x - \mu}{\cancel{\sigma}}\cancel{\sigma}$$
$$z\sigma = x - \mu$$
$$z\sigma + \mu = x - \cancel{\mu} + \cancel{\mu}$$
$$x = z\sigma + \mu$$

Now, you can use this formula to find x when you are given z.

$x = z\sigma + \mu$

$x = 1.43 \cdot 5 + 75$

$x = 7.15 + 75$

$x = 82.15$

Thus, the z-score of 1.43 corresponds to an actual test score of 82.15%.

e. If a student has a z-score of -2.34, what actual score did he get on the test?
$\mu = 75$, $\sigma = 5$, and $z = -2.34$

Use the formula for x from part d of this problem:

$x = z\sigma + \mu$

$x = -2.34 \cdot 5 + 75$

$x = -11.7 + 75$

$x = 63.3$

Thus, the z-score of -2.34 corresponds to an actual test score of 63.3%.

The Five-Number Summary for a Normal Distribution

Looking at the Empirical Rule, 99.7% of all of the data is within three standard deviations of the mean. This means that an approximation for the minimum value in a normal distribution is the mean minus three times the standard deviation, and for the maximum is the mean plus three times the standard deviation. In a normal distribution, the mean and median are the same. Lastly, the first quartile can be approximated by subtracting 0.67448 times the standard deviation from the mean, and the third quartile can be approximated by adding 0.67448 times the standard deviation to the mean. All of these together give the five-number summary.

In mathematical notation, the five-number summary for the normal distribution with mean μ and standard deviation σ is as follows:

Five-Number Summary for a Normal Distribution

$\min = \mu - 3\sigma$

$Q_1 = \mu - 0.67448\sigma$

$\text{med} = \mu$

$Q_3 = \mu + 0.67448\sigma$

$\max = \mu + 3\sigma$

Example 2.4.3: Calculating the Five-Number Summary for a Normal Distribution

Suppose that your class took a test and the mean score was 75% and the standard deviation was 5%. If the test scores follow an approximately normal distribution, find the five-number summary.

The mean is $\mu = 75\%$ and the standard deviation is $\sigma = 5\%$. Thus, the five-number summary for this problem is:

$$\textbf{min} = 75 - 3(5) = 60\%$$
$$Q1 = 75 - 0.67448(5) \approx 71.6\%$$
$$\textbf{med} = 75\%$$
$$Q3 = 75 + 0.67448(5) \approx 78.4\%$$
$$\textbf{max} = 75 + 3(5) = 90\%$$

Section 2.5: Correlation and Causation, Scatter Plots

The label on a can of Planters Cocktail Peanuts says, "Scientific evidence suggest but does not prove that eating 1.5 ounces per day of most nuts, such as peanuts, as part of a diet low in saturated fat and cholesterol & not resulting in increased caloric intake may reduce the risk of heart disease. See nutritional information for fat content (1.5 oz. is about 53 pieces)." Why is it written this way and what does this statement mean? There are many studies that exist that show that two variables are related to one another. The strength of a relationship between two variables is called **correlation**. Variables that are strongly related to each other have strong correlation. However, if two variables are correlated it does not mean that one variable caused the other variable to occur. The above example from the Planters Cocktail Peanuts label is an example of this. There is a strong correlation between eating a diet that is low in saturated fat and cholesterol and heart disease. But that correlation does not mean that eating a diet that is low in saturated fat and cholesterol will cause your risk of heart disease to go down. There could be many different variables that could cause both variables in question to go down or up. One example is that a person's genetic makeup could make them not want to eat fatty food and also not develop heart disease. No matter how strong a correlation is between two variables, you can never know for sure if one variable causes the other variable to occur without conducting experimentation. The only way to find out if eating a diet low in saturated fat and cholesterol actually lowers the risk of heart disease is to do an experiment. This is where you tell one group of people that they have to eat a diet low in saturated fat and cholesterol and another group of people that they have to eat a diet high in saturated fat and cholesterol, and then observe what happens to both groups over the years. You cannot morally do this experiment, so there is no way to prove the statement. That is why the word "may" is in the statement. We see many correlations like this one. Always be sure not to make a correlation statement into a causation statement.

Example 2.5.1: Correlation vs Causation

For each of the following scenarios answer the question and give an example of another variable that could explain the correlation.

a. There is a negative correlation between number of children a woman has and her life expectancy. Does that mean that having children causes a woman to die earlier?

   A correlation between two variables does not mean that one causes the other. A possible cause for both variables could be better health care. If there is better health care, then life expectancy goes up, and also with better health care birth control is more readily available.

b. There is a positive correlation between ice cream sales and the number of drownings at the beach. Does that mean that eating ice cream can cause a person to drown?

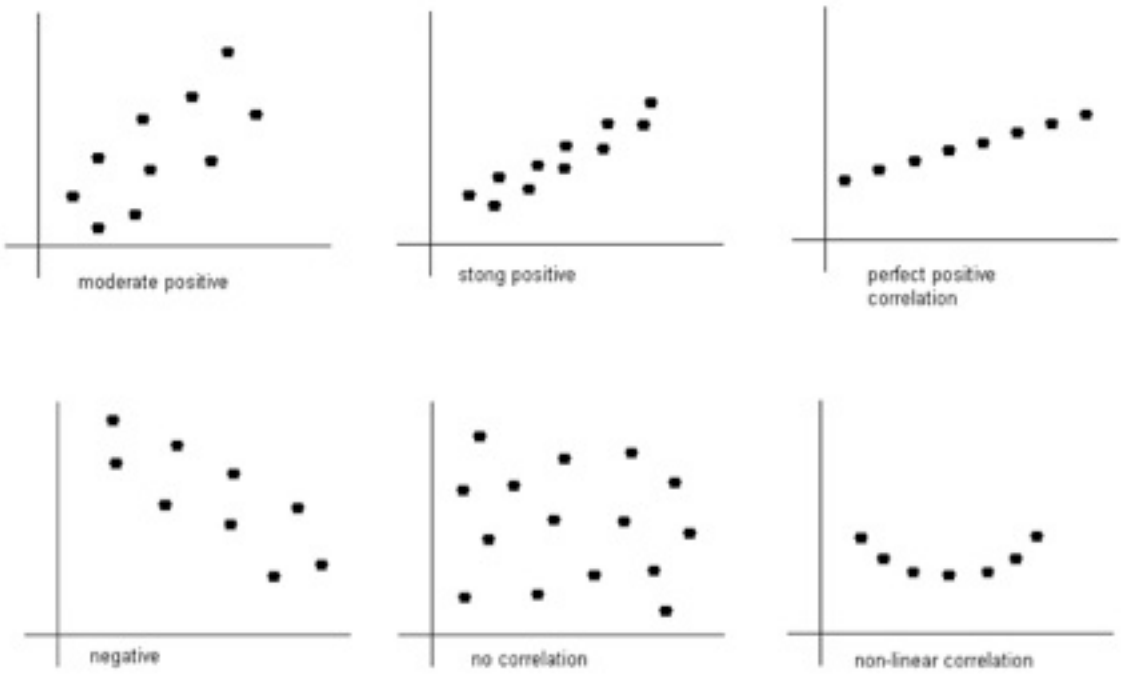
   A correlation between two variables does not mean that one causes the other. The cause for both could be that the temperature is going up. The higher the temperature, the more likely someone will buy ice cream and the more people at the beach.

c. There is a correlation between waist measures and wrist measures. Does this mean that your waist measurement causes your wrist measurement to change?

   A correlation between two variables does not mean that one causes the other. The cause of both could be a person's genetics, eating habits, exercise habits, etc.

How do we tell if there is a correlation between two variables? The easiest way is to graph the two variables together as ordered pairs on a graph called a **scatter plot**. To create a scatter plot, consider that one variable is the independent variable and the other is the dependent variable. This means that the dependent variable depends on the independent variable. We usually set up these two variables as ordered pairs where the independent variable is first and the dependent variable is second. Thus, when graphed, the independent variable is graphed along the horizontal axis and the dependent variable is graphed along the vertical axis. You do not connect the dots after plotting these ordered pairs. Instead look to see if there is a pattern, such as a line, that fits the data well. Here are some examples of scatter plots and how strong the linear correlation is between the two variables.

Figure 2.5.1: Scatter Plots Showing Types of Linear Correlation



Creating a scatter plot is not difficult. Just make sure that you set up your axes with scaling before you start to plot the ordered pairs.

Example 2.5.2: Creating a Scatter Plot

Data has been collected on the life expectancy and the fertility rate in different countries ("World health rankings," 2013). A random sample of 10 countries was taken, and the data is:

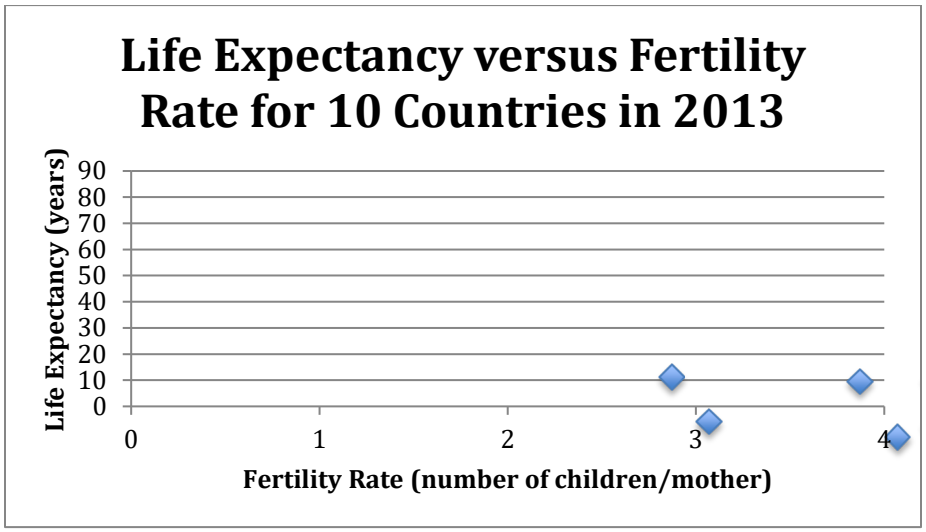
Table 2.5.2: Life Expectancy and Fertility Rate in 2013

Country	Life Expectancy (years)	Fertility Rate (number of children per mother)
Singapore	82.3	1.1
Monaco	81.9	1.8
Canada	81.5	1.6
Ecuador	76	2.5
Malaysia	73.9	3
Lithuania	73.8	1.2
Belize	73.6	3.4
Algeria	73	1.8
Trinidad/tob.	70.8	1.7
Tajikistan	67.9	3

To make the scatter plot, you have to decide which variable is the independent variable and which one is the dependent variable. Sometimes it is obvious which variable is which, and in some case it does not seem to be obvious. In this case, it seems to make more sense to predict what the life expectancy is doing based on fertility rate, so choose life expectancy to be the dependent variable and fertility rate to be the independent variable. The horizontal axis needs to encompass 1.1 to 3.4, so have it range from zero to four, with tick marks every one unit. The vertical axis needs to encompass the numbers 70.8 to 81.9, so have it range from zero to 90, and have tick marks every 10 units.

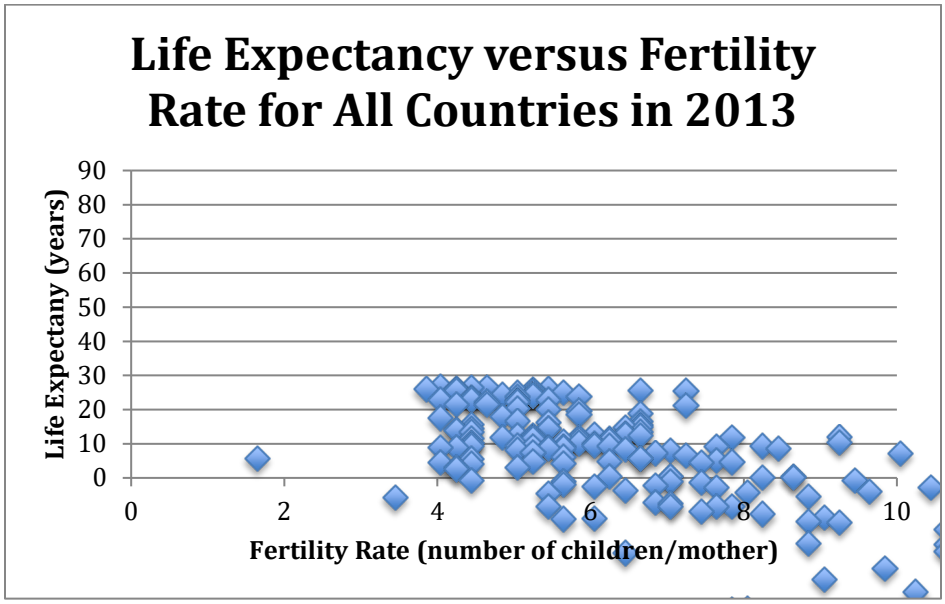
Note: Always start the vertical axis at zero to avoid exaggeration of the data.

Graph 2.5.3: Scatter Plot of Life Expectancy versus Fertility Rate



From the graph, you can see that there is somewhat of a downward trend, but it is not prominent. What this says is that as fertility rate increases, life expectancy decreases. The trend is not strong which could be due to not having enough data or this could represent the actual relationship between these two variables. Let's see what the scatter plot looks like with data from all countries in 2013 ("World health rankings," 2013).

Graph 2.5.4: Scatter Plot of Life Expectancy versus Fertility Rate for All Countries in 2013



Again, there is a downward trend. It looks a little stronger than the previous scatter plot and the trend looks more obvious. This correlation would probably be considered moderate negative correlation. It appears that there is a trend that the higher the fertility rate, the lower the life expectancy. Caution: just because there is a correlation between higher fertility rate and lower life expectancy, do not assume that having fewer children will mean that a person lives longer. The fertility rate does not necessarily cause the life expectancy to change. There are many other factors that could influence both, such as medical care and education. Remember a correlation does not imply causation.

Chapter 2 Homework

1. A study was conducted of Long Beach School District schools regarding how many require school uniforms. In 2006, of the 296 schools questioned, 184 said they required school uniforms. (Gentile & Imberman, 2009) Find the proportion of schools that require a school uniform.

2. A Center for Disease Control (CDC) study conducted in 2008, found that out of 32,601 children in Arizona, 507 had autism. (CDC, 2012) Find the proportion of children in Arizona who had autism in 2008.

3. The temperatures (in degrees Fahrenheit) for the first 10 days of July 2013 in Phoenix, AZ are given in the table below ("Weather underground," 2013).

112	108	111	108	106
111	112	113	107	104

   a. Find the mean, median, and mode for the data set.
   b. Find the range, variance, and standard deviation for the data set.
   c. Find the five-number summary and the interquartile range (IQR) for the data set.
   d. Draw a box-and-whiskers plot for the data set.

4. The number of traffic fatalities involving a driver with a blood alcohol content of 0.1 or more is given in the table for the southern states in 2013 ("Traffic fatalities by," 2013).

260	904	394	366	177	194
264	430	423	345	278	134

   a. Find the mean, median, and mode for the data set.
   b. Find the range, variance, and standard deviation for the data set.
   c. Find the five-number summary and the interquartile range (IQR) for the data set.
   d. Draw a box-and-whiskers plot for the data set.

5. A new sedan car in 2013 or 2014, which has a gas mileage in the city of 40 to 50 mpg, had the following prices ("Motor trend," 2013).

$38,700	$26,200	$39,780	$27,200	$38,700
$24,360	$39,250	$35,925	$35,555	$26,140
$24,995	$28,775	$24,200	$34,755	$25,990

   a. Find the mean, median, and mode for the data set.
   b. Find the range, variance, and standard deviation for the data set.
   c. Find the five-number summary and the interquartile range (IQR) for the data set.
   d. Draw a box-and-whiskers plot for the data set.

6. The prices of an airline flight from New York City to Los Angeles on September 7, 2013 around 8 am, and returning September 14, 2013 are given in the table below ("Expedia," 2013).

$317	$351	$378
$397	$327	$334
$337	$383	$327

   a. Find the mean, median, and mode for the data set.
   b. Find the range, variance, and standard deviation for the data set.
   c. Find the five-number summary and the interquartile range (IQR) for the data set.
   d. Draw a box-and-whiskers plot for the data set.

7. The gas prices (in $/gallon) at all gasoline stations in Flagstaff, AZ, on July 16, 2013 are given in the table below ("Arizona gas prices," 2013).

3.45	3.47	3.48	3.48	3.48	3.49
3.51	3.51	3.51	3.55	3.55	3.56
3.59	3.59	3.59	3.59	3.65	3.65
3.65	3.65	3.66	3.67	3.69	3.69
3.69	3.69	3.69	3.69	3.69	3.69
3.69					

   Using a calculator, find the mean, median, standard deviation, and five-number summary.

8. The city gas mileage (in mpg) of 2011 small pick-up trucks that are four-wheel drive are given in the table below ("Fuel efficiency guide," 2011).

17	18	17	14	16	16
14	14	15	17	18	17
14	16	16	14	14	15
14	18	18	16	14	

Using a calculator, find the mean, median, standard deviation, and five-number summary.

9. Sustainability Victoria, in Australia, surveys all Victorian local communities on waste and recycling services every year. A random sample of 10 local communities reported the number of households served in that community in 2008 and the data is given in the table below ("2001-02 to 2007-08," 2009).

7,551	4,907	45,439	46,000	46,000
49,732	38,264	39,195	40,374	40,500

   a. Find the mean and median of the data set.
   b. Find the mean and the median of the data set with the two lowest data values (7,551 and 4,907) removed.
   c. Discuss what happened to the mean and the median when the two lowest data values (7,551 and 4,907) are removed.

10. Natural gas consumptions (in billions of cubic feet) for selected countries in South America are listed in the following table ("International energy statistics," 2013).

1629	87	885	199
312	8	202	961

   a. Find the mean and median of the data set.
   b. Find the mean and median with the highest data value of 1629 removed.
   c. Discuss what happened to the mean and median when the highest data value (1629) is removed.

11. Suppose your child takes a test to evaluate whether or not your child is at risk for Attention Deficit Hyperactivity Disorder (ADHD). One assessment for ADHD is the Behavior Assessment System for Children, Second Edition (BASC-2) survey. After taking this survey a child is rated on several different qualities. One of the qualities is aggression, where a high score represents a tendency towards being aggressive. Suppose your child is in the 35th percentile on aggression.
   a. What does this percentile mean?
   b. What does this percentile mean about your child and this quality of ADHD?

12. You are planning to go to graduate school after you finish your bachelor's degree and you take the Graduate Record Examination (GRE). Your score on the mathematics section of the general GRE puts you in the 90^{th} percentile.
    a. What does this percentile mean?
    b. Did you pass (score of 70% or better) the mathematics section of the general GRE?

13. The IQ of a person follows a normal distribution and has a mean of 100 and a standard deviation of 15. Using this information, find the following:
    a. What percentage of the people have IQ scores between 85 and 115?
    b. What percentage of the people have IQ scores between 70 and 100?
    c. What percentage of the people have IQ scores between 130 and 145?
    d. What percentage of the people have IQ scores above 145?

14. The mean systolic blood pressure of people in the U.S. is 124 with a standard deviation of 16. Assume that systolic blood pressure follows a normal distribution.
    a. What percentage of the people in the U.S. have systolic blood pressure between 108 and 124?
    b. What percentage of the people in the U.S. have systolic blood pressure between 92 and 156?
    c. What percentage of the people in the U.S. have systolic blood pressure between 76 and 108?
    d. What percentage of the people in the U.S. have systolic blood pressure above 156?

15. The mean diastolic blood pressure of people in the U.S. is 77 with a standard deviation of 11. Assume that diastolic blood pressure follows a normal distribution.
    a. What percentage of the people in the U.S. have diastolic blood pressure between 77 and 88?
    b. What percentage of the people in the U.S. have diastolic blood pressure between 66 and 88?
    c. What percentage of the people in the U.S. have diastolic blood pressure between 55 and 99?
    d. What percentage of the people in the U.S. have diastolic blood pressure below 55

16. The mean height of men in the U.S. is 69.1 inches with a standard deviation of 2.9 inches. Assume that height follows a normal distribution.
    a. What percentage of the males in the U.S. have height between 74.9 and 77.8 inches?
    b. What percentage of the males in the U.S. have height between 60.4 and 77.8 inches?
    c. What percentage of the males in the U.S. have height between 63.3 and 72 inches?
    d. What percentage of the males in the U.S. have heights below 63.3 inches?

17. The IQ of a person follows a normal distribution and has a mean of 100 and a standard deviation of 15. Find the z-score for an IQ score of 134. Is this value unusual? Why or why not?

18. The mean systolic blood pressure of people in the U.S. is 124 with a standard deviation of 16. Assume that systolic blood pressure follows a normal distribution. Find the z-score for a systolic blood pressure of 135. Is this value unusual? Why or why not?

19. The mean diastolic blood pressure of people in the U.S. is 77 with a standard deviation of 11. Assume that diastolic blood pressure follows a normal distribution. Find the z-score for a diastolic blood pressure of 54. Is this value unusual? Why or why not?

20. The mean height of mean in the U.S. is 69.1 inches with a standard deviation of 2.9 inches. Assume that height follows a normal distribution. Find the z-score for a man who is 64 inches tall. Is this value unusual? Why or why not?

21. The IQ of a person follows a normal distribution and has a mean of 100 and a standard deviation of 15. Find the five-number summary.

22. The mean systolic blood pressure of people in the U.S. is 124 with a standard deviation of 16. Assume that systolic blood pressure follows a normal distribution. Find the five-number summary.

23. The mean diastolic blood pressure of people in the U.S. is 77 with a standard deviation of 11. Assume that diastolic blood pressure follows a normal distribution. Find the five-number summary.

24. The mean height of mean in the U.S. is 69.1 inches with a standard deviation of 2.9 inches. Assume that height follows a normal distribution. Find the five-number summary.

25. It can be shown that a man's height and a man's weight have a positive correlation. Does this mean that a man's height causes him to be a certain weight? Explain.

26. Engine size and city gas mileage have a negative correlation. Does this mean that the engine size causes the gas mileage of a car? Explain.

27. Suppose 10 men had their height and weight measured. The data is below. Draw a scatter plot of the data. Describe what relationship you can see from the graph.

Height (inches)	67	72	74	65	70	72	74	69	68	70
Weight (pounds)	185	202	226	165	221	217	218	189	201	185

28. Nine midsize 2011 hybrid cars' city gas mileage and engine size are recorded below ("Fuel efficiency guide," 2011). Draw a scatter plot of the data. Describe what relationship you can see from the graph.

Engine Size	4.4	2.5	2.4	5.0	2.5	2.5	2.5	2.4	1.8
City MPG	17	41	25	19	41	41	33	31	51