CHAPTER **3** # Describing Distributions

Chapter Outline

3.1 Measures of Center

Learning Objectives

- Calculate the mode, median, and mean for a set of data, and understand the differences between each measure of center.
- Identify the symbols and know the formulas for sample and population means.
- Determine the values in a data set that are outliers.

Mode, Mean, and Median

It makes sense to summarize a data set by identifying a value around which the data is centered. Three commonly used statistics that quantify the idea of center are the mode, median and mean. This lesson is an overview of these three basic statistics that are used to measure the center of a set of data.

Mode

The **mode** is defined as the most frequently occurring value in a data set. While many elementary school children learn the mode as their first introduction to measures of center, as you delve deeper into statistics, you will most likely encounter it less frequently. The mode really only has significance for data measured at the most basic of levels. The mode is most useful in situations that involve categorical (qualitative) data that is measured at the nominal level. For example, the mode might be used to describe the most common MM color.

Example A

The students in a statistics class were asked to report the number of children that live in their house (including brothers and sisters temporarily away at college). The data is recorded below:

$$1, 3, 4, 3, 1, 2, 2, 2, 1, 2, 2, 3, 4, 5, 1, 2, 3, 2, 1, 2, 3, 6$$

In this example, the mode could be a useful statistic that would tell us something about the families of statistics students in our school. In this case, 2 is the mode as it is the most frequently occurring number of children in the sample, telling us that a large number of students in our class have two children in their home.

Notice how careful we are to NOT apply this to a larger population and assume that this will be true for any population other than our class! In a later chapter, you will learn how to correctly select a sample that could represent a broader population.

Two Issues with the Mode

a. If there is more than one number that is the most frequent than the mode is both of those numbers. For example, if there were seven 3−child households and seven with 2 children, we would say that the mode is, "2 and 3." When data is described as being **bimodal**, it is clustered about two different modes. Technically, if there were more than two, they would all be the mode. However, the more of them there are, the more trivial

the mode becomes. In those cases, we would most likely search for a different statistic to describe the center of such data.

b. If each data value occurs an equal number of times, we usually say, "There is no mode." Again, this is a case where the mode is not at all useful in helping us to understand the behavior of the data.

Do You Mean the Average?

You are probably comfortable calculating averages. The average is a measure of center that statisticians call the **mean**. Most students learn early on in their studies that you calculate the mean by adding all of the numbers and dividing by the number of numbers. While you are expected to be able to perform this calculation, most real data sets that statisticians deal with are so large that they very rarely calculate a mean by hand. It is much more critical that you understand *why* the mean is such an important measure of center. The mean is actually the numerical "balancing point" of the data set. A certain math teacher might refer to the mean as being "the center of mass".

We can illustrate this physical interpretation of the mean. Below is a graph of the class data from the last example.
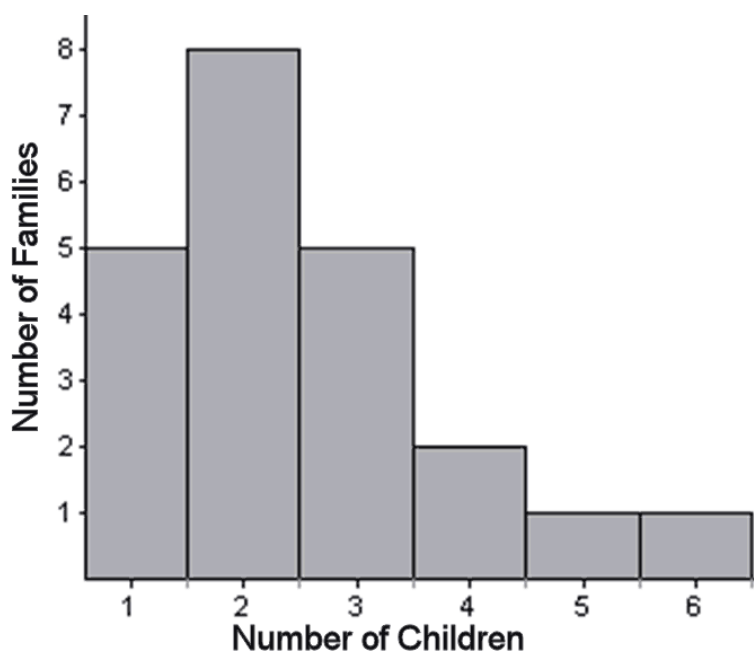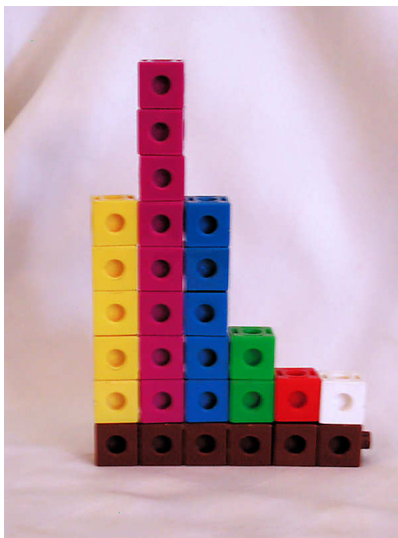

FIGURE 3.1

If you have snap cubes like you used to use in elementary school, you can make a physical model of the graph, using one cube to represent each student's family and a row of six cubes at the bottom to hold them together like this:

There are 22 students in this class and the total number of children in all of their houses is 55, so the mean of this data is $55 \div 22 = 2.5$ children in each student's family. Statisticians use the symbol $\overline{X}$ to represent the mean when X is the symbol for a single measurement. It is pronounced "*x bar*."

It turns out that the model that you created balances at 2.5. In the pictures below, you can see that a block placed at 3 causes the graph to tip left, and while one placed at 2 causes the graph to tip right. However, if you place the block at about 2.5, it balances perfectly!
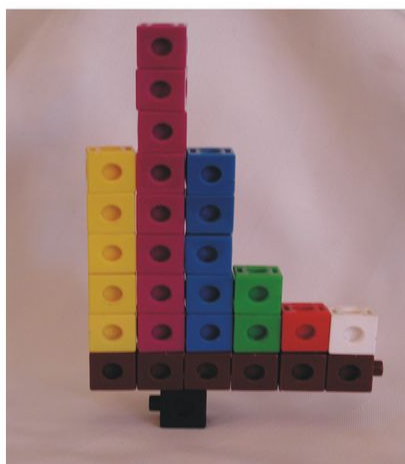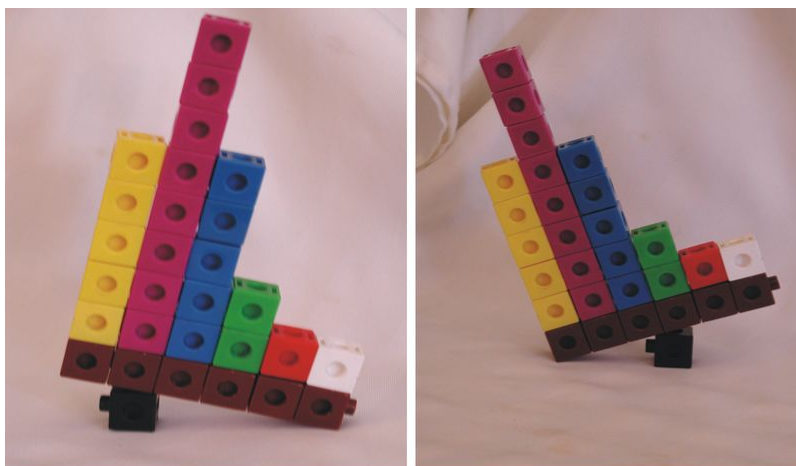
Right Down the Middle: The Median

The **median** is simply the middle number in a set of data. Think of five students seated in a row in statistics class:

Aliyah Bob Catalina David Elaine

Which student is sitting in the middle? If there were only four students, what would be the middle of the row? These are the same issues you face when calculating the numeric middle of a data set using the median.

Let's say that Ron has taken five quizzes in his statistics class and received the following grades:

$$80, 94, 75, 90, 96$$

Before finding the median, you must put the data in order. The median is the numeric middle. Placing the data in order from least to greatest yields:

$$75, 80, 90, 94, 96$$

The middle number in this case is the third grade, or 90, so the median of this data is 90. Notice that just by coincidence, this was also the third quiz that he took, but this will usually not be the case.

Of course, when there is an even number of numbers, there is no true value in the middle. In this case we take the two middle numbers and find their mean. If there are four students sitting in a row, the middle of the row is halfway between the second and third students.

Example B

Take Rhonda's quiz grades:

$$91, 83, 97, 89$$

Place them in numeric order:

$$83, 89, 91, 97$$

The second and third numbers "straddle" the middle of this set. The mean of these two numbers is 90, so the median of the data is 90.

$$\frac{(89 + 91)}{2} = 90 \text{ median}$$

Mean vs. Median

Both the mean and the median are important and widely used measures of center. So you might wonder why we need them both. There is an important difference between them that can be explained by the following example.

Suppose there are 5 houses in a community with the following prices:

> *$55,000 $58,000 $60,000 $61,000 $3,200,000*

The "median housing price" is far more useful in real estate because it tells us that half of the houses on the market are less and half are more. The mean of this data set $686,000, and it is not at all descriptive of the market. The important thing to understand is that extremely large or small values in a data set can have a large influence on the mean.

Shape of a Distribution and Its Relationship to Measures of Center

We typically use the median to describe the center of a skewed distribution. You know how to recognize the mode, or the score with the highest frequency. The mean of a skewed distribution will always be pulled toward the tail of the distribution. And, as shown on the histograms below, the median always falls between the mean and the mode of a skewed distribution.
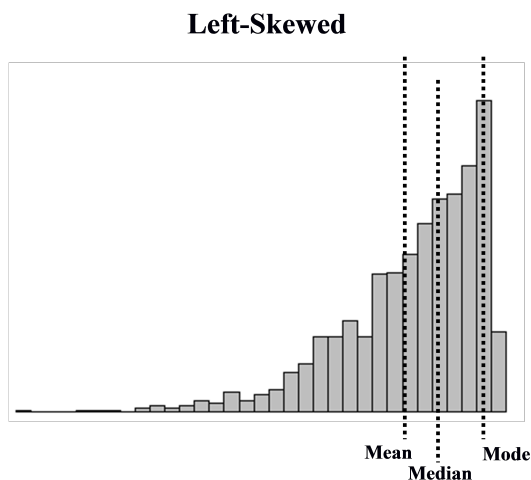
Left-Skewed

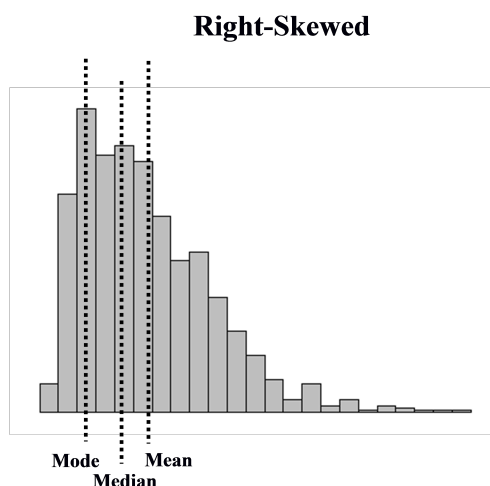FIGURE 3.2

Right-Skewed

FIGURE 3.3

Outliers

So why are the mean and median so different in our earlier example about home values? It is because there is one price that is extremely different from the rest of the data. In statistics, we call such extreme values **outliers**. The mean is affected by the presence of an outlier; however, the median is not. You can see this in the graphs shown above. A statistic that is not affected by outliers is called **resistant**. We say that the median *is* a resistant measure of center, and the mean is not resistant. In a sense, the median is able to *resist* the pull of a far away value, but the mean is drawn to such values. It cannot *resist* the influence of outlier values. Remember the balancing point example? If you created another number that was far away, you would be forced to move the block toward it to make it stay balanced.

Population Mean vs. Sample Mean

Now that we understand some basic concepts about the mean, it is important to be able to represent and understand the mean symbolically. When you are calculating the mean as a statistic from a finite sample of data, we call this the sample mean and as we have already mentioned, the symbol for this is $\overline{X}$. Written symbolically then, the formula for a sample mean is:

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

You may have remembered seeing the symbol $\sum$ before on a calculator or in another mathematics class. It is called "sigma," the Greek capital *S*. In mathematics, we use this symbol as a shortcut for "the sum of". So, the formula is the sum of all the data values (x_1, x_2, etc.) divided by the number of observations (n).

Recall that the mean of an entire population is a parameter. The symbol for a population mean is another Greek letter, μ. It is the lowercase Greek *m* and is called "mu" (pronounced "mew", like the sound a cat makes). In this case the symbolic representation would be:

$$\mu = \frac{\sum(X_1 + X_2 + \cdots + X_n)}{N}$$

The formula is very much the same, because we calculate the mean the same way, but we typically use capital X for the individuals in the population and capital N to represent the size of the population.

In general, statisticians say that $\bar{x}$, the mean of a portion of the population is an estimate of μ, the mean of the population, which is usually unknown. In this course you will learn to determine how good that estimate is.

Lesson Summary

When examining a set of data, we use descriptive statistics to provide information about where the data is centered. The mode is a measure of the most frequently occurring number in a data set and is most useful for categorical data and data measured at the nominal level. The mean and median are two of the most commonly used measures of center. The mean, or average, is the sum of the data points divided by the total number of data points in the set. In a data set that is a sample from a population, the sample mean is notated as $\bar{x}$. When the entire population is involved, the population mean is μ. The median is the numeric middle of a data set. If there are an odd number of numbers, this middle value is easy to find. If there is an even number of data values, however, the median is the mean of the middle two values. The median is resistant, that is, it is not affected by the presence of outliers. An outlier is a number that has an extreme value when compared with most of the data. The mean is not resistant, and therefore the median tends to be a more appropriate measure of center to use in examples that contain outliers. Because the mean is the numerical balancing point for the data, is in an extremely important measure of center that is the basis for many other calculations and processes necessary for making useful conclusions about a set of data.

Points to Consider

a. How do you determine which measure of center best describes a particular data set?
b. What are the effects of outliers on the various measures of spread?
c. How can we represent data visually using the various measures of center?

Review Questions

1. In Lois' second grade class, all of the students are between 45 and 52 inches tall, except one boy, Lucas, who is 62 inches tall. Which of the following statements is true about the heights of all of the students?

   (a) The mean height and the median height are about the same

   (b) The mean height is greater than the median height.

   (c) The mean height is less than the median height.

   (d) More information is needed to answer this question.

   (e) None of the above is true.

2. Enrique has a $91, 87$, and 95 for his statistics grades for the first three quarters. His mean grade for the year must be a 93 in order for him to be exempt from taking the final exam. Assuming grades are rounded following valid mathematical procedures, what is the *lowest* whole number grade he can get for the 4th quarter and still be exempt from taking the exam?
3. The chart below shows the data from the Galapagos tortoise preservation program with just the number of individual tortoises that were bred in captivity and reintroduced into their native habitat.

TABLE 3.1:

Island or Volcano	Number of Individuals Repatriated
Wolf	40
Darwin	0
Alcedo	0
Sierra Negra	286
Cerro Azul	357
Santa Cruz	210
Española	1293
San Cristóbal	55
Santiago	498
Pinzón	552
Pinta	0

For this data, calculate each of the following:

    a. mode
    b. median
    c. mean
    d. explain the difference between your answers to (b) and (c).

Review Answers

1. There is an outlier that is larger than most of the data. This outlier will "pull" the mean towards it while the median tends to stay in the center of the data, clustered somewhere between 45 and 52.
2. His mean for all four quarters would need to be at least 92.5 in order to receive the necessary grade. Multiplying 92.5 by 4, yields 370 as the necessary total. His existing grades total to 273. $370 - 273 = 97$.
3.   a. 0
      b. 210
      c. 299
4. There is one extreme point, 1293, which causes the mean to be greater than the median.

3.2 Range and Interquartile Range

Learning Objectives

- Calculate the range.
- Calculate percentiles and quartiles.
- Calculate the interquartile range.

Introduction

In the previous lesson, we concentrated on statistics that provided information about the way in which a data set is centered. Another important feature that can help us understand more about a data set is the manner in which the data is distributed or **spread**. There are several numbers we can calculate that help us understand how data is spread. This section will focus on the measures of variability (or spread) that can be used to describe distributions of any shape.

Range

For most students, their first introduction to a statistic that measures spread is the **range**. The range is simply the difference between the smallest value (minimum) and the largest value (maximum) in the data. Let's return to the data set used in the previous lesson:

$$75, 80, 90, 94, 96$$

Most students find it intuitive to say that the values range from 75 to 96. However, the range is a statistic, and as such is a single number. It is therefore more proper to say that the range is 21.

The range is useful because it requires very little calculation and therefore gives a quick and easy "snapshot" of how the data is spread, but it is limited because it only involves two values in the data set and it is not resistant to outliers.

Quartiles and Percentiles

A quartile divides the data into four approximately equal groups. The **lower quartile**, sometimes abbreviated as Q_1, is also know as the 25th percentile. A **percentile** is a statistic that identifies the percentage of the data that is *less* than the given value. Technically, the median is a "middle" quartile and is referred to as Q_2. Because it is the numeric middle of the data, half of the data is below the median and half is above. The **upper quartile**, or Q_3, is also know as the 75th percentile.

Your first exposure to percentiles was most likely as a baby. To check a child's physical development, pediatricians use height and weight charts that help them to know how the child compares to children of the same age. A child whose height is in the 70th percentile is taller than 70% of the children of their same age.

Returning to a previous data set:

$$1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 5, 6$$

Recall that the median (50th percentile) of this dataset is 2. The quartiles can be thought of as the medians of the upper and lower halves of the data.
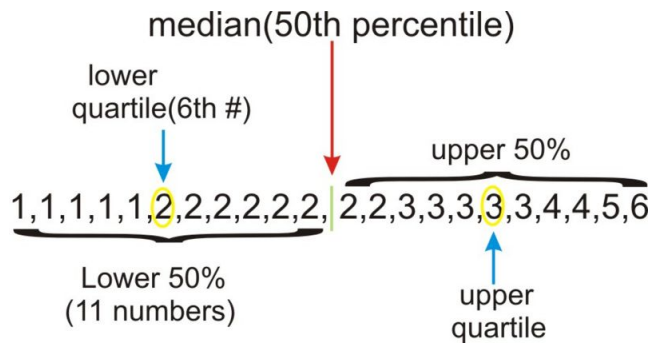
In this case, there are an odd number of numbers in each half. If there were an even number of numbers, then we would follow the procedure for medians and average the middle two numbers of each half. Look at the following set of data:
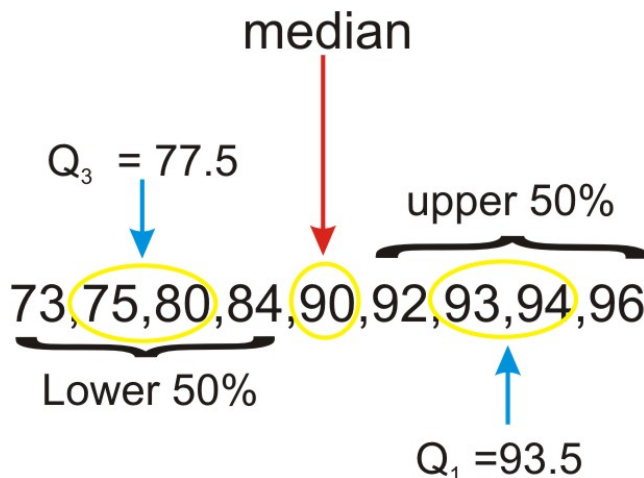
The median in this set is 90. Because it is the middle number, it is not technically part of either the lower or upper halves of the data, so we do not include it when calculating the quartiles. However, not all statisticians agree that this is the proper way to calculate the quartiles in this case. As we mentioned in the last section, some things in statistics are not quite as universally agreed upon as in other branches of mathematics. The exact method for calculating quartiles is another one of those topics.

Interquartile Range

The **interquartile range** (IQR) is the range of the data that contains the middle 50% of cases. Recall that you find the *range* by subtracting the minimum value from the maximum value in the dataset. You calculate in the IQR in a simlar way, except that you find the difference between the 1st quartile (Q_1) and the 3rd quartile (Q_3).

Therefore,

$$IQR = Q3 - Q1$$

Example

A recent study proclaimed Mobile, Alabama the "wettest" city in America. The following table lists a measurement of the approximate annual rainfall in Mobile for the last 10 years. Find the Range and IQR for this data.

TABLE 3.2:

Year	Rainfall (inches)
1998	90
1999	56
2000	60
2001	59
2002	74
2003	76
2004	81
2005	91
2006	47
2007	59

First, place the data in order from smallest to largest. The range is the difference between the minimum and maximum rainfall amounts.

47,56,59,59,60,74,76,81,90,91

RANGE: 91 - 47 = 44

To find the *IQR*, first identify the quartiles, and then subtract $Q3 - Q1$

Q_1 Q_3

47,56,59,59,60|74,76,81,90,91

IQR: 81 - 59 = 22

Even though we are doing easy calculations, statistics is never about meaningless arithmetic and you should always be thinking about what a particular statistical measure means in the real context of the data. In this example, the range tells us that there is a difference of 44 inches of rainfall between the wettest and driest years in Mobile. The *IQR* shows that there is a difference of 22 inches of rainfall even in the middle 50% of the data. It appears that Mobile experiences wide fluctuations in yearly rainfall totals, which might be explained by its position near the Gulf of Mexico and its exposure to tropical storms and hurricanes. The *IQR* will be useful in the next section because it allows us to visual if data is bunched up or spread out.

Points to Consider

    a. How do you determine which measure of center best describes a particular data set?

b. What are the effects of outliers on the various measures of spread?

c. How can we represent data visually using the various measures of center?

Review Questions

1. The chart below shows the data from the Galapagos tortoise preservation program with just the number of individual tortoises that were bred in captivity and reintroduced into their native habitat.

TABLE 3.3:

Island or Volcano	Number of Individuals Repatriated
Wolf	40
Darwin	0
Alcedo	0
Sierra Negra	286
Cerro Azul	357
Santa Cruz	210
Española	1293
San Cristóbal	55
Santiago	498
Pinzón	552
Pinta	0

For this data, calculate each of the following:

a. mode

b. median

c. mean

d. upper and lower quartiles

e. The percentile for the number of Santiago tortoises reintroduced.

Review Answers

1. 270

    a. 0

    b. 210

    c. 222

    d. $Q_1 : 0, Q_3 : 498$

    e. 72.7%

3.3 Five-Number Summary

Learning Objectives

- Calculate the values of the five-number summary.
- Create box-and-whisker plots.
- Interpret the shape of a box-and-whisker plot.

The Five-Number Summary

The **five-number summary** is a numerical description of a data set comprised of the following measures (in order): minimum value, lower quartile, median, upper quartile, maximum value. When you are asked to summarize what you know about a distribution of data, it is often a good starting point to report the five-number summary along with the shape of the distribution.

Example

The huge population growth in the western United States in recent years, along with a trend toward less annual rainfall in many areas and even drought conditions in others, has put tremendous strain on the water resources available now and the need to protect them in the years to come. Here is a listing of the amount of water held by each major reservoir in Arizona stated as a percentage of that reservoir's total capacity.

TABLE 3.4:

Lake/Reservoir	% of Capacity
Salt River System	59
Lake Pleasant	49
Verde River System	33
San Carlos	9
Lyman Reservoir	3
Show Low Lake	51
Lake Havasu	98
Lake Mohave	85
Lake Mead	95
Lake Powell	89

This data set was collected in 1998, and the water levels in many states have taken a dramatic turn for the worse. For example, Lake Powell is currently at less than 50% of capacity[1].

Placing the data in order from smallest to largest gives the following:

$$3, 9, 33, 49, 51, 59, 85, 89, 95, 98$$

Since there are 10 numbers, the median is the average of 51 and 59, which is 55. Recall that the lower quartile is the 25th percentile, or where 25% of the data is below that value. In this data set, that number is 33. Also, the upper quartile is 89. Therefore, the five-number summary is as shown:

$$\{3, 33, 55, 89, 98\}$$

Box-and-Whisker Plots

A **box-and-whisker plot** is a very convenient and informative way to display the information captured in the five number summary. A box-and-whisker plot shows the center and spread of the values on a single quantitative variable. To create the 'box' part of the plot, first draw a rectangle that extends from the lower (first) quartile to the upper (third) quartile. Then draw a line through the interior of the rectangle at the median. Finally, connect the ends of the box to the minimum and maximum values using line segments to form the 'whiskers'.

Here is the box plot for this data:
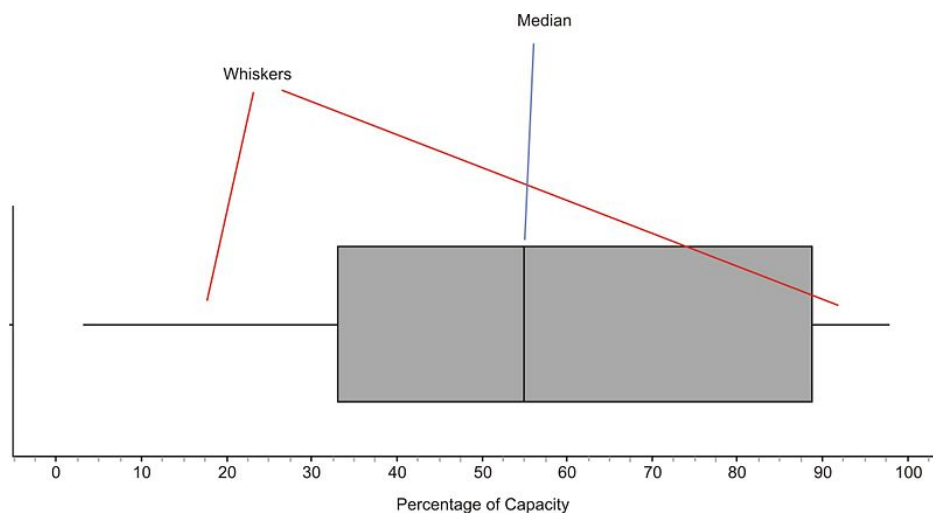
The plot divides the data into quarters. You can also usually learn something about the shape of the distribution from the sections of the plot. If each of the four sections of the plot is about the same length, then the data will be symmetric. In this example, the different sections are not exactly the same length. The left whisker is slightly longer than the right, and the right half of the box is slightly longer than the left. We would most likely say that this distribution is moderately symmetric. In other words, there is roughly the same amount of data in each section. The different lengths of the sections tell us how the data are spread in each section. The numbers in the left whisker (lowest 25% of the data) are spread more widely than those in the right whisker.

Here is the box plot (as the name is sometimes shortened) for reservoirs and lakes in Colorado:
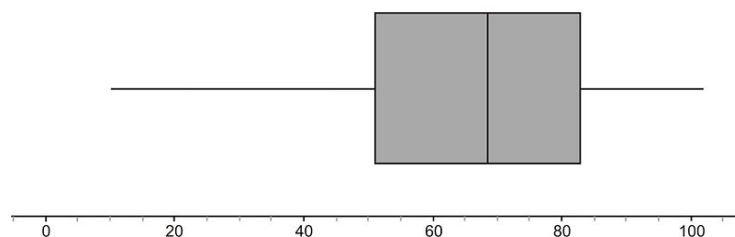

In this case, the third quarter of data (between the median and upper quartile), appears to be a bit more densely concentrated in a smaller area. The data values in the lower whisker also appear to be much more widely spread than in the other sections. Looking at the dot plot for the same data shows that this spread in the lower whisker gives the data a slightly skewed-left appearance (though it is still roughly symmetric).
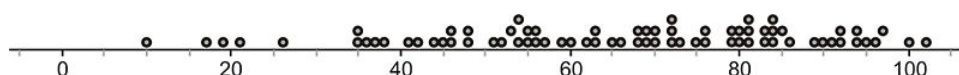
Comparing Multiple Box Plots

Box-and-whisker plots are often used to get a quick and efficient comparison of the general features of multiple data sets. In the previous example, we looked at data for both Arizona and Colorado. How do their reservoir capacities compare? You will often see multiple box plots either stacked on top of each other, or drawn side-by-side for easy comparison. Here are the two box plots:
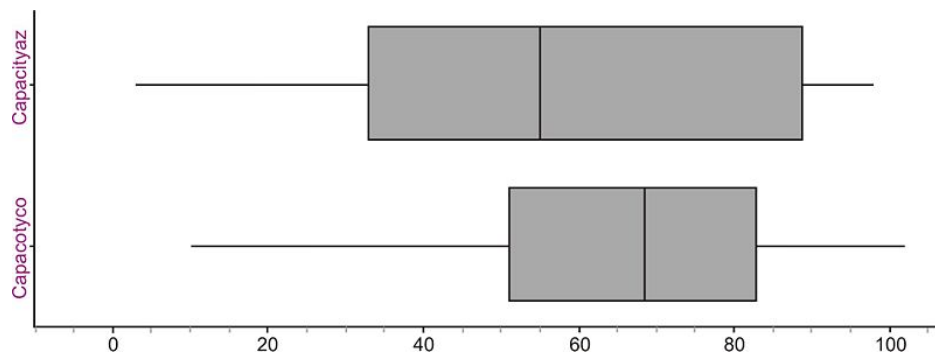
The plots seem to be spread the same if we just look at the range, but with the box plots, we have an additional indicator of spread if we examine the length of the box (or interquartile range). This tells us how the middle 50% of the data is spread, and Arizona's data values appear to have a wider spread. The center of the Colorado data (as evidenced by the location of the median) is higher, which would tend to indicate that, in general, Arizona's reservoirs are less full, as a percentage of their individual capacities, than Colorado's. Recall that the median is a resistant measure of center, because it is not affected by outliers. The mean is not resistant, because it will be pulled toward outlying points. When a data set is skewed strongly in a particular direction, the mean will be pulled in the direction of the skewing, but the median will not be affected. For this reason, the median is a more appropriate measure of center to use for strongly skewed data.

Even though we wouldn't characterize either of these data sets as strongly skewed, this affect is still visible. Here are both distributions with the means plotted for each.

| mean (Capacotyco) = 65.2297
| mean (Capacityaz) = 57.1
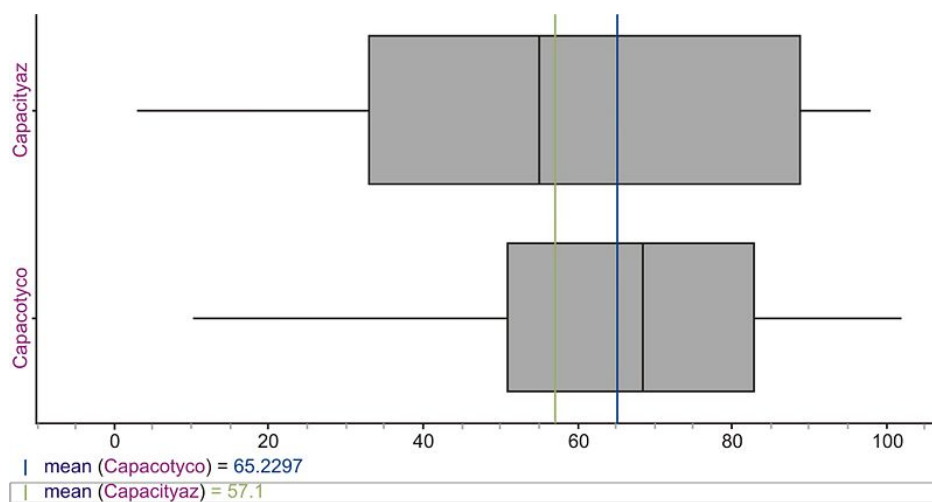
Notice that the long left whisker in the Colorado data causes the mean to be pulled toward the left, making it lower than the median. In the Arizona plot, you can see that the mean is slightly higher than the median, due to the slightly elongated right side of the box. If these data sets were perfectly symmetric, the mean would be equal to the median in each case.

Outliers in Box-and-Whisker Plots

Here are the reservoir data for California (the names of the lakes and reservoirs have been omitted):

$80, 83, 77, 95, 85, 74, 34, 68, 90, 82, 75$

At first glance, the 34 should stand out. It appears as if this point is different from the rest of the data. Notice that without the outlier, the distribution is really roughly symmetric.
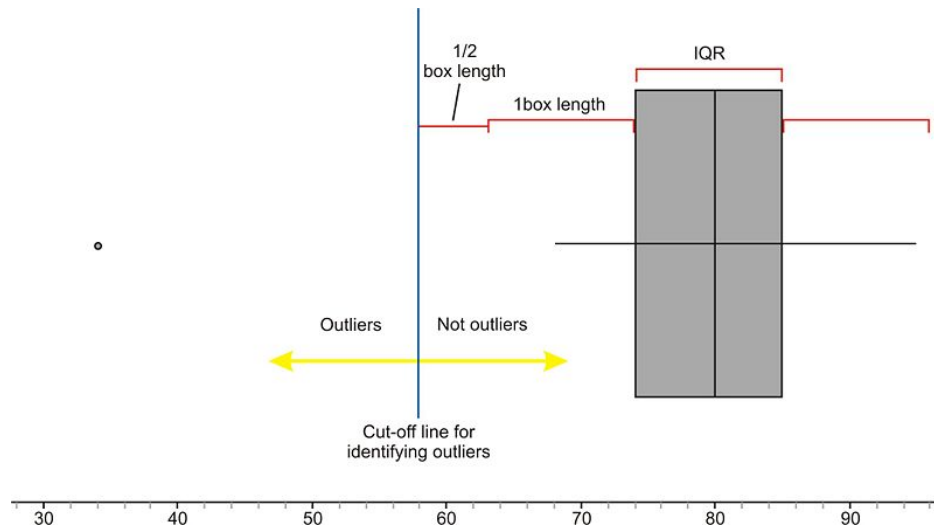
This data set had one obvious outlier, but when is a point far enough away to be called an outlier? We need a standard accepted practice for defining an outlier in a box plot. This rather arbitrary definition is that any point that is more than 1.5 times the *IQR outside the box* will be considered an outlier. Because the *IQR* is the same as the length of the box, any point that is more than one-and-a-half box lengths below Q_1 or above Q_3 is plotted as a separate point and not included in the whisker.

The calculations for determining the outlier in this case are as follows:

Lower Quartile: 74

Upper Quartile: 85

Interquartile range $(IQR): 85 - 74 = 11$

$1.5 * IQR = 16.5$

Cut-off for outliers in left whisker: $74 - 16.5 = 57.5$. Thus, any value less than 57.5 is considered an outlier.

Notice that we did not even bother to test the calculation on the right whisker, because it should be obvious from a quick visual inspection that there are no points that are farther than even one box length away from the upper quartile.

Lesson Summary

The five-number summary is a useful collection of statistical measures consisting of the following in ascending order: minimum, lower quartile, median, upper quartile, maximum. A box-and-whisker plot is a graphical representation of the five-number summary showing a box bounded by the lower and upper quartiles and the median as a line in the box. The whiskers are line segments extended from the quartiles to the minimum and maximum values. Each whisker and section of the box contains approximately 25% of the data. The width of the box is the interquartile range, or *IQR*, and shows the spread of the middle 50% of the data. Box-and-whisker plots are effective at giving an overall impression of the shape, center, and spread of a data set. While an outlier is simply a point that is not

typical of the rest of the data, there is an accepted definition of an outlier in the context of a box-and-whisker plot. Any point that is more than 1.5 times the length of the box (IQR) from either end of the box is considered to be an outlier.

Points to Consider

- What characteristics of a data set make it easier or harder to represent it using dot plots, stem-and-leaf plots, histograms, and box-and-whisker plots?
- Which plots are most useful to interpret the ideas of shape, center, and spread?

Review Questions

1. Here are the 1998 data on the percentage of capacity of reservoirs in Idaho.

$$70, 84, 62, 80, 75, 95, 69, 48, 76, 70, 45, 83, 58, 75, 85, 70$$
$$62, 64, 39, 68, 67, 35, 55, 93, 51, 67, 86, 58, 49, 47, 42, 75$$

1.
   a. Find the five-number summary for this data set.
   b. Show all work to determine if there are true outliers according to the $1.5 * IQR$ rule.
   c. Create a box-and-whisker plot showing any outliers.
   d. Describe the shape, center, and spread of the distribution of reservoir capacities in Idaho in 1998.
   e. Based on your answer in part (d), how would you expect the mean to compare to the median? Calculate the mean to verify your expectation.

2. Here are the 1998 data on the percentage of capacity of reservoirs in Utah.

$$80, 46, 83, 75, 83, 90, 90, 72, 77, 4, 83, 105, 63, 87, 73, 84, 0, 70, 65, 96, 89, 78, 99, 104, 83, 81$$

2.
   a. Find the five-number summary for this data set.
   b. Show all work to determine if there are true outliers according to the $1.5 * IQR$ rule.
   c. Create a box-and-whisker plot showing any outliers.
   d. Describe the shape, center, and spread of the distribution of reservoir capacities in Utah in 1998.
   e. Based on your answer in part (d) how would you expect the mean to compare to the median? Calculate the mean to verify your expectation.
3. Graph the box plots for Idaho and Utah on the same axes. Write a few statements comparing the water levels in Idaho and Utah by discussing the shape, center, and spread of the distributions.
4. If the median of a distribution is less than the mean, which of the following statements is the most correct?

   (a) The distribution is skewed left.

   (b) The distribution is skewed right.

   (c) There are outliers on the left side.

   (d) There are outliers on the right side.

   (e) Answers (B) or (D) could be true.

5. The following table contains recent data on the average price of a gallon of gasoline for states that share a border crossing into Canada.

    a. Find the five-number summary for this data.
    b. Show all work to test for outliers.
    c. Graph the box-and-whisker plot for this data

TABLE 3.5:

State	Average Price of a Gallon of Gasoline (US$)
Alaska	3.458
Washington	3.528
Idaho	3.26
Montana	3.22
North Dakota	3.282
Minnesota	3.12
Michigan	3.352
New York	3.393
Vermont	3.252
New Hampshire	3.152
Maine	3.309

3.4 Interpreting Box-and-Whisker Plots

Learning Objective

- Interpret box-and-whisker plots.
- Answer questions about the data by referencing the box-and-whisker plot for the distribution.

Brief Review

Box-and-whisker plots (or *"box plots"*) are commonly used to compare a single value or range of values for easier, more effective decision-making. Box and whisker plots are very effective and easy to read, and can summarize data from multiple sources and display the results in a single graph.

Use box and whisker plots when you have multiple data sets from independent sources that are related to each other in some way. Examples include comparing test scores between schools or classrooms, and exploring data from before and after a process change.

Remember that the line inside the box represents the middle value when the data points are arranged numerically. Because the median is only identified by location in a series, it can sometimes be very indicative of the trend or average of the data set as a whole, and sometimes is not useful for that purpose at all (see Example A).

Recall that skewed data appears as a longer "tail" in one direction on a histogram, it is similar on a box plot. If the box in a box plot is stretched in one direction or the other, then the data is skewed in that direction. Data skewed right indicates a closer concentration of values on the *left,* since the plot indicates values more "strung out" on the right side.

A longer box indicates a greater **interquartile range** since the sides of the box indicate the 1st and 3rd quartiles. A greater interquartile range is an indicator of data that may be somewhat unreliable. Since the interquartile range represents the 50% of the data closest to the median, a greater range in this section of the plot suggests that the median may not be a great indicator of central tendency.

A plot with long whiskers represents a greater range for the overall sample than simply a longer box itself does. Data covering a greater range is naturally less reliable as an indicator of highly probable values, but given the option, longer whiskers are less of a concern than a long box. A broad range of possibilities but a strong likelihood of central values is more reliable to use for prediction than a moderate overall range with little concentration at the median.

Example A

Identify the 5 number summary and any outliers depicted in the box plot below:

Solution

The 5 number summary is depicted by the vertical bars in the box and by the endpoints of the 'whiskers':

- Minimum: 13
- 1st Quartile: 16
- Median: 19
- 3rd Quartile: 22
- Maximum: 24
- Outliers (depicted by open circles disconnected from the box and whiskers): 4 and 30

Example B

What is indicated by the shape of the box plot below?
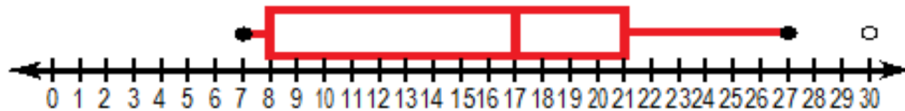
Solution

The box in the plot extends nearly to the lower extreme, indicating that the data less than the median is likely at least relatively consistent, since there is not a large jump between the lower 25% and the minimum. The longer whisker on the upper side suggests that there may be larger variance among the greater values, since there is a greater distance from the 3rd quartile to the upper extreme than from the median to the 3rd quartile.

Lesson Summary

If you were asked to evaluate a box plot to find the median, quartiles, extremes and outliers, would you know how? What does it mean if the 'box' in a box plot is unusually long or short? Does a long 'whisker' on one or both sides mean something important?

With the practice you have had now, these questions should be easy!

- Median: the center vertical line in the 'box'
- 1st and 3rd Quartiles: the leftmost and rightmost vertical lines of the 'box'
- Lower and Upper Extremes: the endpoints of the 'whiskers'

Vocabulary

The **interquartile range** is calculated by subtracting the 1st quartile from the 3rd quartile and represents the middle 50% of the sample.

Guided Practice

1. Make a Box and Whisker plot from the following data sets.

    (a) Initial weight (December) of 14 women in a weight loss study (pounds)$190, 175, 187, 199, 205, 187, 176, 180, 187, 191$

1. (b) Weights of the same women one month later (January) $187, 174, 181, 189, 196, 178, 174, 176, 181, 186, 188, 191, 183, 1$

1. (c) Weights of the same women in February $181, 165, 176, 182, 190, 176, 171, 170, 171, 185, 187, 181, 179, 186$

2. How do the data in (a) and (c) compare?
3. How did the median change?
4. How did the maximum weight change?
5. How did the minimum weight change?
6. How did the range change?
7. How would you judge the effectiveness of the weight loss method used in the study?

Solutions

1. For all three sets, first organize the data by increasing numerical order and identify the five-number summary (FNS). Once you have the FNS, create the box plot for each just as in the examples above. The three plots should resemble the images below:


(a)

(b)

(c)
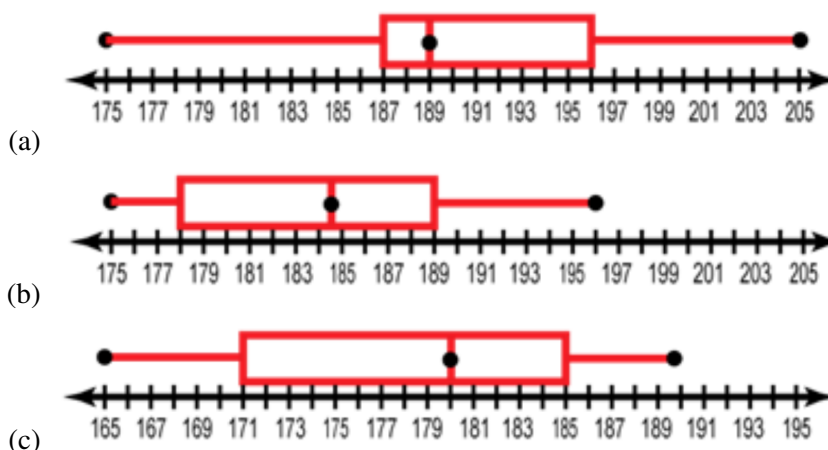
2. If we compare the boxplots for (a) and (c), we can see that the median weight has dropped by about 9 pounds. In addition, there are essentially no individuals in boxplot (c) that weigh above the median in boxplot (a).
3. The median in December was 189, and in February it was 180.
4. The maximum in December was 205, and went down to 190 by February.
5. The minimum weight in December was 175, and it also went down, to 165 by February.
6. The range decreased notably, from 30 pounds in December, to 25 pounds in February.
7. It would appear that the method was effective, at least in the short term.

More Practice

1. What is the five number summary of the following box and whisker plot?
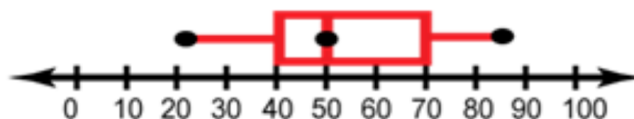
2. The box plot shows the heights in inches of boys on a High School Baseball Team. What is the 5 number summary of the plot?
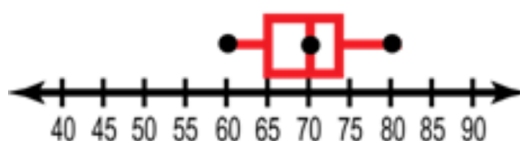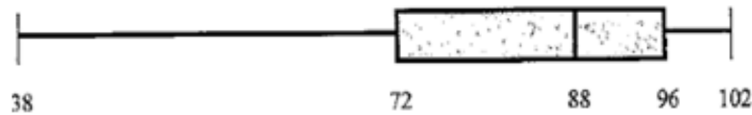
3. Listed are the heights in inches of girls on a High School Ski Team. Make a plot of the girls' heights.

 $58, 59, 59, 60, 62, 65, 68, 69, 70, 70, 71$

4. Comparing the heights between the two teams, which has the taller players on average? How do you know?

Use the box and whisker plot below to examine scores received on an English GED Test to answer questions 5-9.

5. What was the highest score on the test?
6. What percent of the class scored above a 72?
7. What was the median score on the test?
8. What percent of the class scored between 88 and 96?
9. Would you expect the mean to be above or below the median? Explain.

**Use the graph below that shows how much girls spent on average per month on clothes during August.**
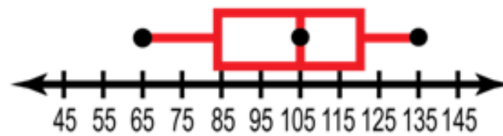
10. How many girls shop for clothes? (Hint: can you answer this question?)
11. What percent of girls spent less than $85.00 in August on clothes?
12. Would you expect the mean number of dollars spent to be higher or lower than the median? Explain.

Use the graphs below to compare the amount of time a teenager spends in the bathroom getting ready for school and the amount of time they spend in the bathroom getting ready to go to a party.

TIME SPENT GETTING READY FOR SCHOOL
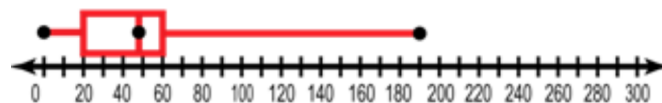
TIME SPENT GETTING READY FOR A PARTY
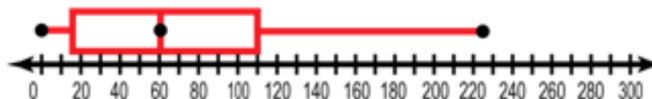
13. What percent of teenagers spend at least 15 minutes getting ready for a party?
14. What is the 3rd Quartile for the time spent getting ready for a party?
15. Is it more common for a teenager to spend more than 1 hour getting ready for school or between 1 and 2 hrs getting ready for a party? Explain.

Answer True or False for questions 16-24.

16. _____ Some teenagers do not spend time getting ready for parties.
17. _____ The graph of time spent getting ready for a party contains more data than the getting ready for school graph.
18. _____ 25% of teenagers spend between 48 and 60 minutes getting ready for school.
19. _____ 15% of the teenagers did not go to parties that month
20. _____ In general teenagers spend more time getting ready for a party than getting ready for school.
21. _____ The Party data is more varied than the getting ready for school data
22. _____ The ratio of teenagers who spend more than 110 minutes getting ready for a party to those who spend less is about 2:1
23. _____ 225 Teenagers watch TV.
24. _____ Twice as many teenagers spend more than 1 hour on getting ready for school, than they do spending an hour getting ready for a party.