3 | NUMERICAL DESCRIPTORS

Figure 4.1 Photo by Dayne Topkin on Unsplash

Introduction

Chapter Objectives

By the end of this chapter, the student should be able to:

- Recognize, describe, and calculate the measures of central tendencies: mean, median, and mode.
- Recognize, describe, and calculate the measures of variation: range, variance, and standard deviation.
- Recognize, describe, and calculate the measures of position: quartiles, percentiles, outliers, Chebyshev's Theorem, and Empirical Rule

3.1 | Measures of Central Tendency

The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of a data set are the **mean** (average) and the **median**. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

NOTE

The words “mean” and “average” are often used interchangeably. The substitution of one word for the other is common practice. The technical term is “arithmetic mean” and “average” is technically a center location. However, in practice among non-statisticians, “average" is commonly accepted for “arithmetic mean.”

The symbol used to represent the **sample mean** is an x with a bar over it (pronounced “x bar”): $\bar{x}$.

The Greek letter m (pronounced "mew") represents the **population mean**. One of the requirements for the **sample mean** to be a good estimate of the **population mean** is for the sample used to be truly random.

When the values in a data set are not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. To see that both ways of calculating the mean are the same, consider the sample: 1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4

$$\bar{x} = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7$$

$$\bar{x} = \frac{3(1)+2(2)+1(3)+5(4)}{11} = 2.7$$

You can quickly find the location of the median by using the expression $\frac{n+1}{2}$.

The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered. For example, if the total number of data values is 97, then

$\frac{n+1}{2} = \frac{97+1}{2} = 49,$ so the median is the 49th value in the ordered data. If the total number of data values is 100, then $\frac{n+1}{2} = \frac{100+1}{2} = 50.5,$ so the median occurs midway between the 50th and 51st values.

In general, the values of the mean and median are **not** the same. The upper-case letter *M* is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Example 3.1

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):

3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24;
24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47

Calculate the mean and the median.

Solution 3.1

The calculation for the mean is:

$$\bar{x} = \frac{3+4+(8)(2)+10+11+12+13+14+(15)(2)+(16)(2)+...+35+37+40+(44)(2)+47}{40} = 23.6$$

To find the median, *M*, first use the formula for the location. The location is $\frac{n+1}{2} = \frac{40+1}{2} = 20.5.$

Starting at the smallest value, the median is located between the 20th and 21st values (the two 24s); so

$M = \frac{24+24}{2} = 24.$

Using the TI-83, 83+, 84, 84+ Calculator

To find the mean and the median:

Go to STAT >> EDIT. Clear list L1 by pressing 4: ClrList, and then pressing 2nd 1 for list L1. Press ENTER. Enter data into the list editor. Go to STAT >> EDIT and enter the data values into list L1.

Press STAT and arrow to CALC. Select 1:1-VarStats. Press 2nd 1 for L1 and then ENTER. At the top of the screen you will see $\bar{x}$ = 23.6. Use the arrow keys to scroll down to the second output screen, and you will see the median: Med = 24.

Try It Σ

3.1 The following data show the number of months patients typically wait on a transplant list before getting surgery. The data are ordered from smallest to largest. Calculate the mean and median.

3; 4; 5; 7; 7; 7; 7; 8; 8; 9; 9; 10; 10; 10; 10; 10; 11;
12; 12; 13; 14; 14; 15; 15; 17; 17; 18; 19; 19; 19; 21 ; 21; 22; 22
23; 24; 24; 24; 24

Example 3.2

Suppose that in a small town of 50 people, one person earns $5,000,000 per year and the other 49 each earn $30,000. Which is the better measure of the "center": the mean or the median?

Solution 3.2

The mean is $\bar{x} = \frac{5{,}000{,}000 + 49(30{,}000)}{50} = 129{,}400$, whereas the median is M = 30,000.

The median is a better measure of the "center" than the mean. The mean is distorted because of one very large value, 5,000,000. Such a value is called an “outlier”, meaning that it is an extreme value. The 30,000 gives us a better sense of the middle of the data.

Try It Σ

3.2 In a sample of 60 households, one house is worth $2,500,000. Half of the rest are worth $280,000, and all the others are worth $315,000. Which is the better measure of the “center”: the mean or the median?

Another measure of the center is the mode. The **mode** is the most frequent value. There can be more than one mode in a data set as long as those values have the same frequency, and that frequency is the highest. A data set with two modes is called **bimodal.** There also can be **no mode**. No mode exists when each data value has the same frequency.

Example 3.3

Statistics exam scores for 20 students are as follows:

50; 53; 59; 59; 63; 63; 72; 72; 72; 72; 72; 76; 78; 81; 83; 84; 84; 84; 90; 93

Find the mode.

Solution 3.3

The most frequent score is 72, which occurs five times. Mode = 72.

Try It

3.3 The number of books checked out from the library from 25 students are as follows:

0; 0; 0; 1; 2; 3; 3; 4; 4; 5; 5; 7; 7; 7; 7; 8; 8; 8; 9; 10; 10; 11; 11; 12; 12

Find the mode.

Example 3.4

When is the mode the best measure of the "center"?

Solution 3.4

Consider a weight loss program that advertises a mean weight loss of six pounds the first week of the program. The mode might indicate that most people lose two pounds the first week, making the program less appealing.

NOTE

The mode can be calculated for qualitative data as well as for quantitative data. For example, if the data set is: {red, red, red, green, green, yellow, purple, black, and blue}, then the mode is red.

Statistical software will easily calculate the mean, the median, and the mode. Some graphing calculators can also make these calculations. In the real world, people make these calculations using software.

Try It Σ

3.4 Five credit scores are 680, 680, 700, 720, 720. The data set is bimodal because the scores 680 and 720 each occur twice. Consider the annual earnings of workers at a factory. The mode is $25,000 and occurs 150 times out of 301. The median is $50,000 and the mean is $47,500. What would be the best measure of the "center" for each of these situations?

The Law of Large Numbers and the Mean

The Law of Large Numbers says that if you take samples of larger and larger size from any population, then the mean of the sample is highly likely to get closer and closer to m. Similarly, if we take larger and larger samples, the sample proportion will approach the population proportion p. These facts are discussed in more detail later in the text. Recall that a **statistic** is a number calculated from a sample, whereas a **parameter** is a corresponding number calculated from a population. Examples of statistics include the mean, median, mode and sample proportion.

The Law of Large Numbers tells that $\bar{x}$ is a sample statistic that estimates the parameter m (the population mean).

Similarly, the sample proportion $\hat{p}$ is a sample statistic that estimates the population proportion, p.

Calculating the Mean from Frequency Tables

Suppose that thirty randomly selected students were asked the number of movies they watched the previous week. The results are shown in the frequency table below.

# of movies	Frequency
0	5
1	15
2	6
3	3
4	1

Table 3.1

Suppose that we wanted to know the mean number of movies watched for the sample. Of course, this will be the sum of the data, divided by 30 (the size of the sample):

$$\text{sample mean} = \frac{\text{data sum}}{\text{number of data values}}.$$

To do this, we add up the data values, multiplied by their respective frequencies and divide by 30:

$$\bar{x} = \frac{0(5) + 1(15) + 2(6) + 3(3) + 4(1)}{30} \approx 1.33$$

Note that this could also be written as:

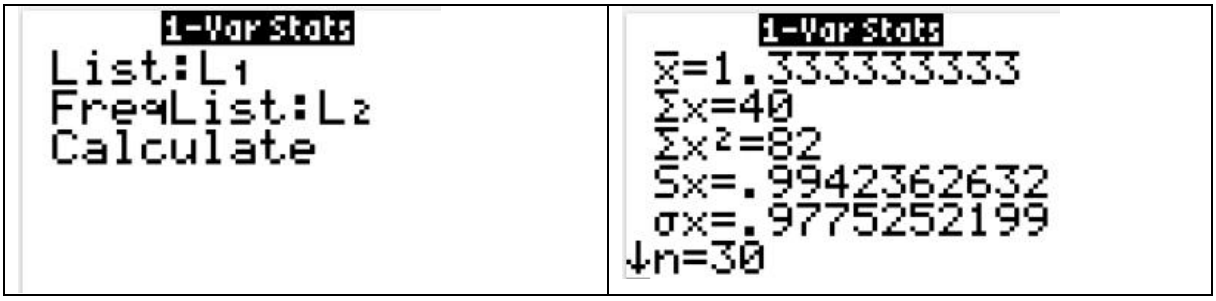
$$\bar{x} = \frac{0(5) + 1(15) + 2(6) + 3(3) + 4(1)}{30} = \frac{0 \cdot 5}{30} + \frac{1 \cdot 15}{30} + \frac{2 \cdot 6}{30} + \frac{3 \cdot 3}{30} + \frac{4 \cdot 1}{30}$$
$$= 0 \cdot \frac{1}{30} + 1 \cdot \frac{15}{30} + 2 \cdot \frac{6}{30} + 3 \cdot \frac{3}{30} + 4 \cdot \frac{1}{30} \approx 1.33$$

In other words, we can also calculate the mean by multiplying the data values by their respective relative frequencies and adding the resulting values. And this can be easily done using the TI-84 family of calculators:

Using the TI-83, 83+, 84, 84+ Calculator

To calculate the mean of a data set from a frequency table

Go to STAT >> EDIT and clear lists L1 and L2.
Enter the data values into L1 and enter the frequencies into L2. Then go to STAT >> CALC. Select 1-VarStats, type L1, L2 and hit ENTER. The mean will appear at the top of the screen as $\bar{x}$.



When only grouped data is available, we do not know the individual data values (we only know intervals and interval frequencies); therefore, we cannot compute an exact mean for the data set. However, we can still estimate the sample mean by from the frequency table. To calculate the mean from a grouped frequency table we can apply the same method as above; but since we do not know the individual data values, we instead use the midpoint of each interval. The midpoint is the average of the lower boundary and upper boundary:

$$\text{midpoint} = \frac{\text{upper boundary} + \text{lower boundary}}{2}.$$

Example 3.5

A frequency table displaying Professor Blount's last statistic test is shown below:

Grade Interval	Number of Students
50-56.5	1
56.5-62.5	0
62.5-68.5	4
68.5-74.5	4
74.5-80.5	2
80.5-86.5	3
86.5-92.5	4
92.5-98.5	1

Find the best estimate of the class mean.

Solution 3.5

First, find the midpoints for all intervals:

Grade Interval	Midpoint	Number of Students
50.5-56.5	53.5	1
56.5-62.5	59.5	0
62.5-68.5	65.5	4
68.5-74.5	71.5	4
74.5-80.5	77.5	2
80.5-86.5	83.5	3
86.5-92.5	89.5	4
92.5-98.5	95.5	1

- Go to STAT >> EDIT, clear lists L1, L2. Enter the midpoints into L1 and the frequencies into L2.
- Go to STAT >> CALC, select 1-VarStats and type L1, L2 (don't forget the comma for the TI-83!).
- Hit ENTER, and the mean will appear at the top of the output screen: $\overline{x} = 76.86$.
- For TI-84 (plus) type L1 into FREQ: and L2 into FREQLIST:, then CALCULATE

Try It Σ

3.5 Maris conducted a study on the effect that playing video games has on memory recall. As part of her study, she compiled the following data:

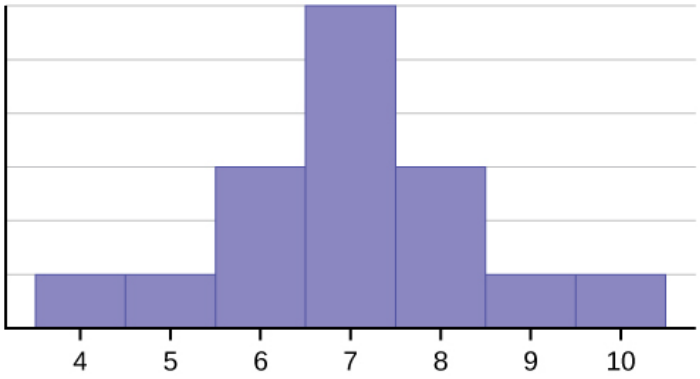
Hours Spent on Video Games	Number of teenagers
0 – 3.5	3
3.5 – 7.5	7
7.5 – 11.5	12
11.5 – 15.5	7
15.5 – 19.5	9

What is the best estimate for the mean number of hours spent playing video games?

Skewness and the Mean, Median, and Mode

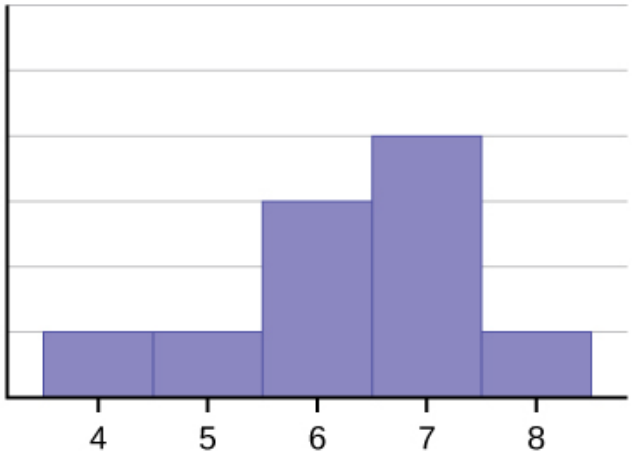
Consider the following data set: 4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10

This data set can be represented by following histogram. Each interval has width one, and each value is in the middle of an interval.



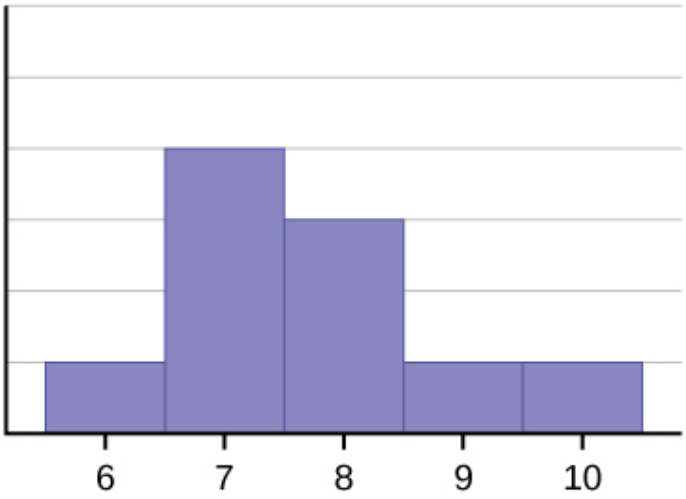
The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each seven for these data. **In a perfectly symmetrical distribution, the mean and the median are the same.** This example has one mode (unimodal), and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

On the other hand, the histogram for the data: 4; 5; 6; 6; 6; 7; 7; 7; 7; 8 is not symmetrical. The right-hand side seems "chopped off" compared to the left side. A distribution of this type is called **skewed to the left** because it is pulled out to the left.



The mean is 6.3, the median is 6.5, and the mode is seven. **Notice that the mean is less than the median, and they are both less than the mode.** The mean and the median both reflect the skewing, but the mean reflects it more so.

Finally, consider the histogram for the data: 6; 7; 7; 7; 7; 8; 8; 8; 9; 10.
This graph also is not symmetrical. It is **skewed to the right**.



The mean is 7.7, the median is 7.5, and the mode is seven. Of the three statistics, **the mean is the largest, while the mode is the smallest**. Again, the mean reflects the skewing the most.

To summarize:

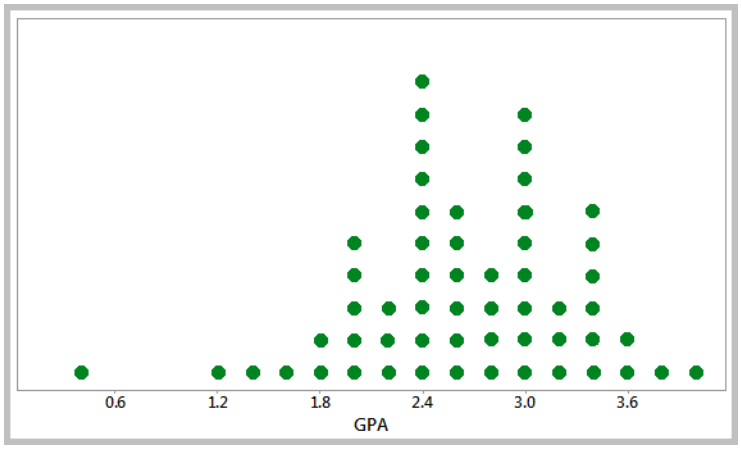
- If the distribution of data is skewed to the left, the mean is less than the median.
- If the data is symmetrical, then the median and mean are equal.
- If the data is skewed to the right, then the mean is greater than the median.

Skewness and symmetry become important when we discuss probability distributions in later chapters.

Try It Σ

3.6 Discuss the mean, median, and mode for the following two graphs (dot plot and histogram).

The GPA's of all students in a business college. Each dot represents up to 4 observations.



3.2 | Measures of Variation

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. In this section we will discuss three measures of variation: the range, the standard deviation, and the variance.

Standard deviation.

The most used measure of variation, or spread, is the standard deviation. The **standard deviation** is a non-negative number that measures how far data values are from their mean. Its importance will become much clearer when we begin studying methods of Inferential Statistics. For now, we point out two key features of this important statistic.

The standard deviation provides a measure of the overall variation in a data set. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket *A* and supermarket *B*. the average wait time at both supermarkets is five minutes. At supermarket *A*, the standard deviation for the wait time is two minutes; at supermarket *B* the standard deviation for the wait time is four minutes. Because supermarket B has a higher standard deviation, we know that there is more variation in the wait times at supermarket *B*. Overall, wait times at supermarket *B* are more spread out from the average; wait times at supermarket *A* are more concentrated near the average, and hence more predictable.

The standard deviation can be used to determine whether a data value is close to or far from the mean. Suppose that Rosa and Binh both shop at supermarket *A*. Rosa waits at the checkout counter for seven minutes and Binh waits for one minute. At supermarket *A*, the mean waiting time is five minutes and the standard deviation is two minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa's wait time of seven minutes is two minutes longer than the average of five minutes; that is, her wait time is **one standard deviation above the average** of five minutes.

Binh's wait time is four minutes less than the average of five minute; that is, his wait time of one minute is **two standard deviations below the average** of five minutes. A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be "unusual", or "far" from the average. Considering data to be far from the mean if it is more than two standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the

data affects how much of the data is further away than two standard deviations. (As we will see in the next section)

As with measures of center, we can have a standard deviation measured from sample data (a statistic) or from population data (a parameter). The lower-case letter *s* represents a sample standard deviation and the Greek letter σ (sigma, lower case) represents a population standard deviation. This is like our notation protocol for the mean, where we used the symbol $\overline{x}$ for the sample mean and the Greek letter m for the population mean.

Calculating the Standard Deviation

If *x* is a data value, then the difference between *x* and the mean is called its **deviation**. In a data set, there are as many deviations as there are values in the data set. The deviations are used to calculate the standard deviation; in fact, we intuitively think of the standard deviation as a sort of "average" of the deviations. If the data belong to a population, in symbols a deviation is $x - \mu$. For sample data, a deviation will be $x - \overline{x}$.

The formula for calculating the standard deviation will also be slightly different for sample data than for population data. If the sample is reasonably large and representative of the population, then *s* should be a good estimate of σ.

Suppose that we wanted to measure the totality of all variation in a data set. Then we might start by adding *all* of the deviations $x - \overline{x}$. However, in any data set, there are values both above and below the mean $\overline{x}$. So, some deviations would be positive, and some will be negative, and there will be considerable cancellation. In fact, the sum of all the deviations would always be zero. To prevent deviations from cancelling each other out we look at the *squares* of the deviations: terms of the form

$$(x - \overline{x})^2.$$

The **variance is the average of the squares of the deviations.** For reasons that will become clear later, we compute these averages slightly differently for population data than we do for sample data.

The population variance is: $\sigma^2 = \sum \frac{(x-\mu)^2}{N}$,

where N = the number of data values in the population.

The sample variance is: $s^2 = \sum \frac{(x-\overline{x})^2}{n-1}$,

where n = the number of data values in the sample.

The symbol σ^2 represents the population variance - so **the population standard deviation σ is the square root of the population variance**. Similarly, the symbol s^2 represents the sample variance, so **the sample standard deviation *s* is the square root of the sample variance**. Taking the square root of the variance can be thought of as reversing the effect of squaring the deviations. Thus, we can think of the standard deviation as a sort of "average" of the deviations. Note also that the standard deviation will always be measured in the same units as the data values.

Formulas for the Standard Deviation:

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

Population Standard Deviation

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

Sample Standard Deviation

NOTE

In practice, we will always use a calculator or computer software to calculate a standard deviation. If you are using a TI-83, 83+, 84+ calculator (or Excel) you need to select the appropriate standard deviation σ_x or S_x from the summary statistics.

We will concentrate on using and interpreting the information that the standard deviation gives us. However, it is instructive to do one step-by-step example to help you understand how the standard deviation measures variation from the mean. The calculator instructions appear at the end of this example.

Example 3.6

In a fifth-grade class, teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a sample of $n = 20$ fifth grade students. The ages are rounded to the nearest half year:

9; 9.5; 9.5; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5;

The average age is 10.53 years, rounded to two places. Find the sample standard deviation, rounded to two decimal places.

Solution 3.6

The variance may be calculated by using a table. Then the standard deviation is obtained by taking the square root of the variance. We will explain the parts of the table after calculating *s*.

Data	**Frequency**	**Deviation**	**Deviation^2**	**Freq * Deviation^2**
x	**f**	**(x – x bar)**	**(x – x bar)^2**	**f * (x – x bar)^2**
9	1	9 – 10.525 = -1.525	$(-1.525)^2$ = 2.325625	1 * 2.325625 = 2.325625
9.5	2	9.5 – 10.525 = -1.025	$(-1.025)^2$ = 1.050625	2* 1.050625 = 2.101250
10	4	10 – 10.525 = -0.525	$(-0.525)^2$ = 0.275625	4 * 0.275625 = 1.1025
10.5	4	10.5 – 10.525 = -0.025	$(-0.025)^2$ = 0.000625	4 * 0.000625 = 0.0025
11	6	11 – 10.525 = 0.475	$(0.475)^2$ = 0.225625	6 * 0.225625 = 1.35375
11.5	3	11.5 – 10.525 = 0.975	$(0.975)^2$ = 0.950625	3 * 0.950625 = 2.851875
				Total = 9.7375

The sample variance, s^2, is equal to the sum of the last column (9.7375) divided by the total number of data values minus one, which is 19.

$$s^2 = \frac{9.7375}{19} = 0.5125$$

The **sample standard deviation** *s* is equal to the square root of the sample variance:

$$s = \sqrt{0.5125} = 0.715891$$

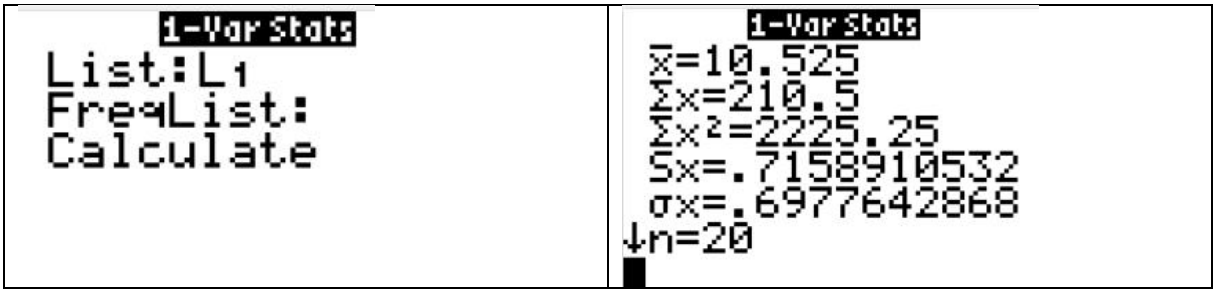
Rounded to two decimal places, ***s* = 0.72**.

Again, we almost never use the formulas or the procedure above to calculate a standard deviation. Instead, we will use either software or a calculator to calculate these.

Using the TI-83, 83+, 84, 84+ Calculator

To find the standard deviation:

Go to STAT >> EDIT. Clear list L1, and enter the data into the list.
Go to STAT >> CALC and select 1:1-VarStats. Press 2nd 1 for L1 and then ENTER.



In the output screen you will see both Sx and σx:

If the data is from a sample, use **Sx** as it represents the ***sample*** **standard deviation, s**.

If the data is from a population use σ_x represents the ***population*** **standard deviation, σ**.

Try It

3.7 On a baseball team, the ages of each of the players are as follows: 21; 21; 22; 23; 24; 24; 25; 25; 28; 29; 29; 31; 32; 33; 33; 34; 35; 36; 36; 36; 36; 38; 38; 38; 40

Use a calculator or computer to find the mean and standard deviation. Then find the value that is two standard deviations above the mean.

Standard deviation of Grouped Frequency Tables

Recall that for grouped data we do not know individual data values, so we cannot describe the typical value of the data with precision. Just as we could not find the exact mean from a frequency table, neither can we find the exact standard deviation. However, by again replacing the class intervals by their midpoints, we can estimate the standard deviation using the same procedure as we used for estimating the mean.

Example 3.7

Find the standard deviation for the data in the following table:

Class	Frequency
0-2	1
3-5	6
6-8	10
9-11	7
12-14	0
15-17	2

Solution 3.7

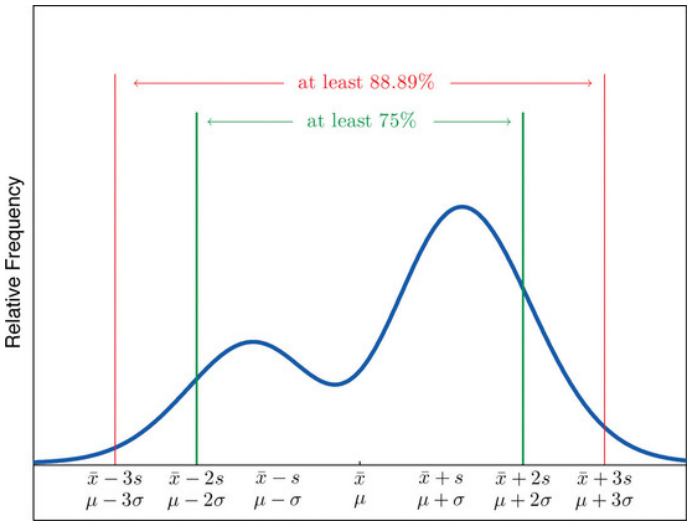
- Go to STAT >> EDIT, and clear lists L1 and L2.
- Enter the midpoints of the classes into L1; enter the frequencies into L2:
- Next, go to STAT >> CALC and select **1-VarStats:**
- Finally, type L1, L2 and hit Enter to get:

Here we see both a population standard deviation, σ_x, and the sample standard deviation S_x displayed.

The following rules provide a little more insight into what the standard deviation tells us about the distribution of the data.

Chebyshev's Rule: For **any** data set, no matter what the distribution of the data is:

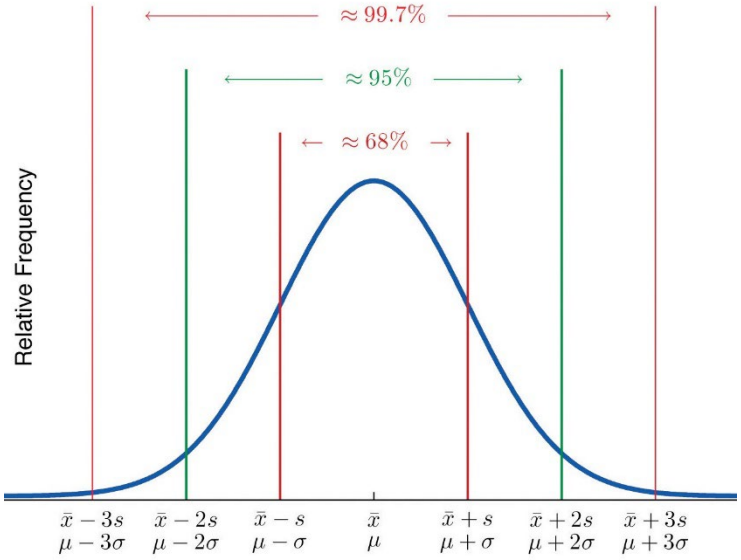
- At least 75% of the data is within 2 standard deviations of the mean.
- At least 89% of the data is within 3 standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.



Chebyshev's Theorem

Empirical Rule: For data having a distribution that is **bell-shaped** and **symmetric**:

- 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.
- It is important to note that this rule applies *only* when the shape of the distribution of the data is bell-shaped and symmetric. We will learn more about this when studying "Normal" distribution in later chapters.



Empirical Rule

Finally, we have seen how the mean and standard deviation can be affected by extreme high or extreme low values. An **outlier** is a data value that is far removed from the rest of the data. Outliers need to be examined carefully - sometimes they are the result of measurement error and should be removed from the data and not used in the analysis at all. Other times, they hold valuable information about the population under study and should be included in the data. Now that we have introduced the standard deviation, we can present a commonly used criterion for classifying a data value as an outlier**:**

A data value that is more than three standard deviations from the mean is called an outlier.

Example 3.8

Using Chebyshev's theorem and Empirical find the percentage of values that lie within 3 standard deviations from the mean; given the heights of 20-year-old males have a symmetrical distribution with a mean of 70.2 inches and a standard deviation of 1.5 inches.

Solution 3.8:

We are given m = 70.2 and s = 1.5. Three standard deviations is represented by 3×s = 3(1.5) = 4.5. Subtract and add 4.5 from the mean, 70.2.

$$70.2 - 4.5 = 65.7$$

$$70.2 + 4.5 = 74.7$$

Chebyshev's theorem states that 89% of 20-year-old male heights lie within 65.7 inches and 74.7 inches. While Empirical Rule states that 99.7% of 20-year-old male heights lie within 65.7 inches and 74.7 inches. Chebyshev's theorem is more conservative since it is used for any type of distribution. Empirical Rule is specifically for symmetrical distributions.

3.3 | Measures of Relative Position

Measures of relative position are used to describe data values relative to other data values. Here we will discuss commonly used measures of relative position – in particular, quartiles, percentiles, and z-scores.

Quartiles

Quartiles are special cases of percentiles. The first quartile, Q_1, is the same as the 25th percentile, and the third quartile, Q_3, is the same as the 75th percentile. The median M is both the second quartile and the 50th percentile. Thus 25% of the data values in any data set are less than Q_1. Similarly, 75% of the data values in the data set are less Q_3, and so 25% of data values are *above* Q_3. Finally, 50% of all data values are between Q_1 and Q_3.

The quartiles separate the data into quarters. To find the quartiles, first find the median or second quartile. The first quartile, Q_1, is the middle value of the lower half of the data, and the third quartile, Q_3, is the middle value, or median, of the upper half of the data. To get the idea, consider the following data set:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

- There are 14 data values, so the median is the average of the 7th and 8th values in the list. Thus, we have $M = (7.2 + 6.8)/2 = 7$.
- The **first quartile** will be the median of the lower half of the data: 1, 1, 2, 2, 4, 6, 6.8. The middle value of these seven data values is $Q_1 = 2$.
- The **third quartile** will be the median of the upper half of the data: 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of these seven data values is $Q_3 = 9$.

The **interquartile range** is the difference between the third quartile (Q_3) and the first quartile (Q_1):

$$IQR = Q_3 - Q_1$$

The IQR provides a numerical measure of the overall amount of variation in a data set; it is often used in conjunction with the median. When the median is used to describe the center, the IQR is often used to describe the spread.

The IQR can also be used to determine potential **outliers**. Recall that an outlier is a data value that is significantly separated from the other data. Using the IQR, we have the following criteria for classifying a data value as an outlier**:**

A data value x is considered an outlier if it is either x < Q1 – 1.5(IQR) or x > Q3 + 1.5(IQR).

Example 3.9

The following data represent selling prices for homes in a certain city. Calculate the *IQR* and determine if any prices are potential outliers. Prices are in dollars.

389,950; 230,500; 158,000; 479,000; 639,000; 114,950; 5,500,000;
387,000; 659,000; 529,000; 575,000; 488,800; 1,095,000

Solution 3.9 Order the data from smallest to largest:

114,950; 158,000; 230,500; 387,000; 389,950; 479,000; 488,800;
529,000; 575,000; 639,000; 659,000; 1,095,000; 5,500,000

There are 13 data values, so the median is the 7th value: $M = 488{,}800$

Q_1 will be the median of the lower six values: $Q_1 = (230{,}500 + 387{,}000)/2 = 308{,}750$.
Q_3 will be the median of the upper six values: $Q_3 = (639{,}000 + 659{,}000)/2 = 649{,}000$.

Thus, the inter-quartile range is $IQR = 649{,}000 - 308{,}750 = 340{,}250$

$\Rightarrow Q1 - (1.5)(IQR) = 308{,}750 - (1.5)(340{,}250) = -201{,}625$

$\Rightarrow Q3 + (1.5)(IQR) = 649{,}000 + (1.5)(340{,}250) = 1{,}159{,}375$

No house price is less than –201,625. However, 5,500,000 is more than 1,159,375. Therefore, the data value 5,500,000 is a potential **outlier**.

Try It

3.8 Two classes take a test. The test Scores for Class *A* are:
69; 96; 81; 79; 65; 76; 83; 99; 89; 67; 90; 77; 85; 98; 66; 91; 77; 69; 80; 94

The test Scores for Class *B* are:
90; 72; 80; 92; 90; 97; 92; 75; 79; 68; 70; 80; 99; 95; 78; 73; 71; 68; 95; 100

Find the interquartile range for the following two data sets and compare them. Which class had more variable test scores?

Five Number Summary and Boxplots

The quartiles are part of the **five-number summary** - these five numbers consist of:

The minimum value
The first quartile Q_1
The median M
The third quartile Q_3
The maximum value

For example, for the real estate data in Example 3.9, the five-number summary would be:

Min = 114,950
Q_1 = 308,750
M = 488,800
Q_3 = 649,000
Max = 5,500,000

For a small example like this, the quartiles can be easily calculated by hand; but for a large data set this can be tedious and time-consuming. But most statistical software packages and graphing calculators will calculate the five-number summary for us. When entering the data into the calculator or software packages, the data does *not* have to be in order.

Using the TI-83, 83+, 84, 84+ Calculator

To find the Five Number Summary:

- Go to STAT >> EDIT. Clear list L1 and enter the data into the list.
- Go to STAT >> CALC and select 1:1-VarStats. Press 2nd 1 for L1 and then ENTER.
- Use the down arrow key to scroll down; the 5-number summary appears in the 2nd output screen.

A **box plot** is a graphical representation of the five-number summary. These graphs (also called **box and whisker plots** or **box-whisker plots**) allow us to see where provide an image of the concentration of the data and show how far the extreme values are from most of the data. So, the box plot gives a good, quick picture of the data.

A box plot is constructed from the five values: the minimum value, the first quartile, the median, the third quartile, and the maximum value. To construct a box plot, use a horizontal number line and mark the five numbers along the axis. We make small vertical marks above the minimum and maximum values, and larger vertical marks above the quartiles Q_1, M, and Q_3. Finally, we "box off" the marks above the quartiles.

So, the first quartile marks one end of the box and the third quartile marks the other end of the box. And the **middle 50 percent of the data fall inside the box.** The "whiskers" extend from the ends of the box to the smallest and largest data values.

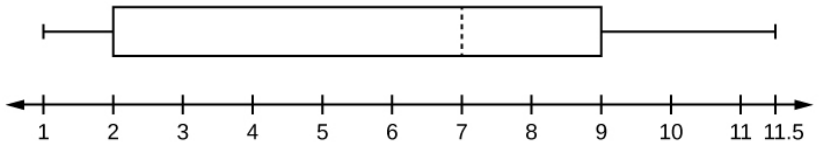
As an example, consider again the dataset:

1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5.

We calculated the quartiles already, and know that the five number summary is:

Min = 1; $Q_1 = 2$; $M = 7$; $Q_3 = 9$; Max = 11.5

Using this information, the boxplot is:



NOTES

1. It is important to start a box plot with a **scaled number line**. Otherwise, the box plot may distort the spread of the data.

2. You may encounter box-and-whisker plots that have dots marking outlier values. In those cases, the whiskers are not extending to the minimum and maximum values.

Example 3.10

The following data are the heights (in inches) of 40 students in a statistics class:

59; 60; 61; 62; 62; 63; 63; 64; 64; 64;
65; 65; 65; 65; 65; 65; 65; 65; 65; 66;
66; 67; 67; 68; 68; 69; 70; 70; 70; 70;
70; 71; 71; 72; 72; 73; 74; 74; 75; 77

a) Find the five-number summary for the data and construct a boxplot.
b) Find the *IQR*
c) Given an interval that has the middle 50% of the data.

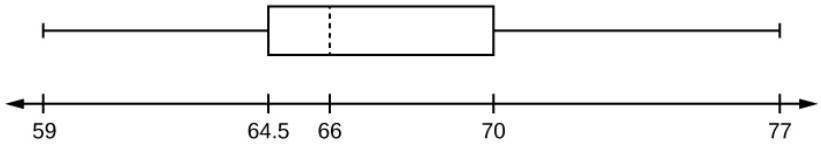
Solution 3.10:

a.) To find the five-number summary, we use the TI-84 calculator.

Go to STAT >> EDIT, and enter the data into L1. Then go to STAT >> CALC and select 1-VarStats. Press 2nd and 1 to select L1, and hit ENTER. Scroll down to the second output screen to get:

Min = 59, Q_1 = 64.5, Q_2: = M = 66, Q_3 = 70, Max = 77

Using these values, we get the following graph:



b) The interquartile range is $IQR = Q_3 - Q_1 = 70 - 64.5 = 5.5$ inches.

c) About 50% of the data will be between Q_1 and Q_3. That is, about 50% of the students in the class had heights that were between 64.5 inches and 70 inches.

Finally, note that the calculator can also be used to construct the box plot.

Using the TI-83, 83+, 84, 84+ Calculator

To construct a boxplot:

- Go to the STAT PLOT menu by pressing 2nd and then the Y = button.
- Turn on Plot 1 (Plot 2, Plot 3 should be turned off).
- Use the down arrow key to move to Type, and select the boxplot option (middle of second row).
- For Xlist, type 2nd and then 1 to select L1.
- Press Zoom and then press 9 to select the ZoomStat graphing option; this will automatically configure the window so that all of the data are included.

Finally, if we press TRACE, and use the arrow keys to move from left to right, we can see the quartiles.

Try It Σ

3.9 The following data are the number of pages in 40 books on a shelf. Construct a box plot using a graphing calculator and state the interquartile range.

136; 140; 178; 190; 205; 215; 217; 218; 232; 234;
240; 255; 270; 275; 290; 301; 303; 315; 317; 318;
326; 333; 343; 349; 360; 369; 377; 388; 391; 392;
398; 400; 402; 405; 408; 422; 429; 450; 475; 512

It is possible that in a data set, two or more numbers in the five-number summary are equal. For instance, you might have a data set in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median.

As an example, suppose we have data set in which the smallest value and the first quartile were both 1, the median and the third quartile were both 5, and the maximum value is 7; then the box plot would look like:

In this case, at least 25% of the values are equal to 1. Twenty-five percent of the values are between 1 and 5, inclusive. At least 25% of the values are equal to 5, and the top 25% of the values fall between 5 and 7, inclusive.

Example 3.11

Given the following test scores solve the following questions:

Test scores for a college statistics class held during the day are:
99; 56; 78; 55.5; 32; 90; 80; 81; 56; 59; 45; 77; 84.5; 84; 70; 72; 68; 32; 79; 90

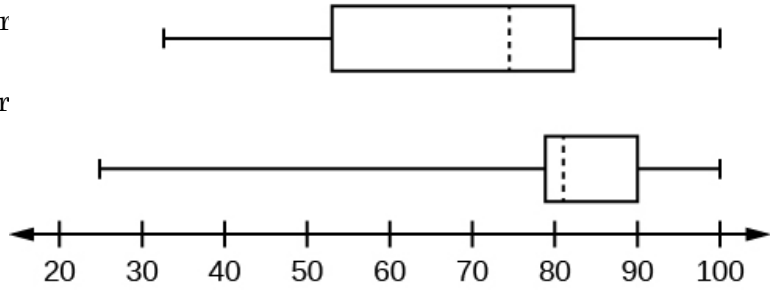
Test scores for a college statistics class held during the evening are:
98; 78; 68; 83; 81; 89; 88; 76; 65; 45; 98; 90; 80; 84.5; 85; 79; 78; 98; 90; 79; 81; 25.5

a. Find the five-number summary for the day class.
b. Find the five-number summary for the night class.
c. Create a box plot for each set of data, using the same number line for both box plots.
d. Which box plot has the widest spread for the middle 50% of the data?
What does this mean for that set of data in comparison to the other set of data?

Solution 3.11

a. Min = 32, Q1 = 56, M = 74.5, Q3 = 82.5, Max = 99
b. Min = 25.5, Q1 = 78, M = 81, Q3 = 89, Max = 98

c. The upper graph is for day class.
The lower graph is for night class.



d. The first data set has the wider spread for the middle 50% of the data; so the *IQR* for the first data set is greater than the *IQR* for the second set. This means that there is more variability in the middle 50% of the first data set.

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order from smallest to largest. For example, 15% of data values are less than or equal to the 15th percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is "good" or "bad". The interpretation will depend on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good;" in other contexts a high percentile might be considered "good". In many situations, there is no value judgment that applies. Understanding how to interpret percentiles properly is important not only when describing data, but also when calculating probabilities in later chapters of this text.

GUIDELINES

When interpreting a percentile in the context of the given data, the sentence should contain the following information:

- Information about the context of the situation being considered
- the data value (value of the variable) that represents the percentile
- the percent of individuals or items with data values below the percentile
- the percent of individuals or items with data values above the percentile.

Example 3.12

On a timed math test, the first quartile for time it took to finish the exam was 35 minutes. Interpret the first quartile in the context of this situation.

Solution 3.12:

- Twenty-five percent of students finished the exam in 35 minutes or less.
- Seventy-five percent of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

Try It Σ

3.10 For the 100-meter dash, the third quartile for times for finishing the race was 11.5 seconds. Interpret the third quartile in the context of the situation.

Example 3.13

On a 20 question math test, the 70th percentile for number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

Solution to 3.13:

- Seventy percent of students answered 16 or fewer questions correctly.
- Thirty percent of students answered 16 or more questions correctly.
- A higher percentile could be considered good, as answering more questions correctly is desirable.

Try It Σ

3.11 On a 60 point written assignment, the 80th percentile for the number of points earned was 49. Interpret the 80th percentile in the context of this situation.

Try It Σ

3.12 During a college basketball season, the 40th percentile for points scored per player in a game is eight. Interpret the 40th percentile in the context of this situation.

Percentiles

We say that a data value x is at the P-th percentile if P% of all data values are less than x. For example, if a student taking a standardized test scores at the 85th percentile, this means that 85% of all students taking the test scored lower than the student, and so 15% of students taking the test scored better.

Percentiles divide ordered data into hundredths and are useful for comparing values. For example, universities and colleges use percentiles extensively. One instance in which colleges and universities use percentiles is when SAT results are used to determine a minimum testing score that will be used as a criterion for admission. For example, suppose Duke accepts SAT scores at or above the 75th percentile. That translates into a score of at least 1220.

Percentiles are most often used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant. We will learn more about how to calculate percentiles for large data sets in a later chapter. But sometimes we also need to calculate a percentile for small data sets. The best way to do this is using a cumulative relative frequency table.

Example 3.14

Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results are as follows:

Hours of Sleep	Frequency	Relative Frequency	Cumulative Relative Frequency
4	2	0.04	0.04
5	5	0.10	0.14
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

a. Find the 28th percentile.
b. Find the median.
c. Find the third quartile.

Solution 3.14:

a. **Find the 28th percentile**. Notice the 0.28 in the "cumulative relative frequency" column. Twenty-eight percent of the 50 is a total of 14 data values. There are 14 values less than the 28th percentile. They include the two 4's, the five 5's, and the seven 6's. The 28th percentile is between the last six and the first seven. **The 28th percentile is 6.5.**

b. **Find the median**. Look again at the "cumulative relative frequency" column and find 0.52. The median is the 50th percentile. Since 50% of 50 is 25, there are 25 data values less than the median. They include the two 4s, the five 5's, the seven 6's, and eleven of the 7's. The median is between the 25th and 26th data values, both of which are 7. The median is $M = 7$.

c. **Find the third quartile**. The third quartile is the same as the 75th percentile. You can "eyeball" this answer. If you look at the "cumulative relative frequency" column, you find 0.52 and 0.80. When you have all the 4's, 5's, 6's and 7's, you have 52% of the data. When you include all the 8's, you will have 80% of the data. **The 75th percentile must be 8**. Another way to look at the problem is to find 75% of 50, which is 37.5, and round up to 38. The third quartile, Q_3, is the 38th value, which is again an 8.

A Formula for Finding the k^{th} Percentile

In addition to using a frequency distribution, there are several formulas for calculating the kth percentile. Here is one of them.

Suppose that we want the k^{th} percentile. It may or may not be part of the data. Order the data, and let i be the index of the k^{th} percentile; i.e. this is the position that the percentile appears in the list of data. Let n be the total number of data values. Then we calculate

$$i = \frac{k}{100}(n+1)$$

If i is a positive integer, then the k^{th} percentile is the data value in the i^{th} position in the ordered set of data. If i is not a positive integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set. This is easier to understand in an example.

Example 3.15

Listed below are the ages for 29 Academy Award winning best actors:

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47;
52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the 70th percentile.
b. Find the 83rd percentile.

Solution 3.15

a. $k = 70$, $i =$ the index, $n = 29$. Calculate $i = \frac{k}{100}(n+1) = \frac{70}{100}(29+1) = 21$.

This is a whole number, so we want the data value in the 21st position in the ordered data. The 70th percentile is **64**.

b. $k = 83$, $i =$ the index, $n = 29$. Calculate $i = \frac{k}{100}(n+1) = \frac{83}{100}(29+1) = 24.9$.

This is *not* a whole number, so we round down to get 24, and round up to get 25.
The age in the 24th position is 71, and age in the 25th position is 72. Average these two numbers to obtain the 83rd percentile, **71.5.**

Try It Σ

3.13 Listed below are the ages for 29 Academy Award winning best actors:

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47;
52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

Calculate the 20th percentile and the 55th percentile.

A Formula for Finding the Percentile of a Value in a Data Set

Again, order the data from smallest to largest.

Let x = the number of data values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile.

Let y = the number of data values equal to the data value for which you want to find the percentile.

Let n = the total number of data values.

Calculate $\frac{x + 0.5y}{n} \cdot 100$ and round to the nearest integer.

Example 3.16

Listed below are the ages for 29 Academy Award winning best actors:

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47;
52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find percentile for 58
b. Find percentile for 25

Solution 3.16

a. Counting from the bottom of the list, there are 18 data values less than 58; and there is one data value equal to 58. So $x = 18$ and $y = 1$.

$$\Rightarrow \frac{x + 0.5y}{n} \cdot 100 = \frac{18 + 0.5(1)}{29} \cdot 100 = 63.8$$

Thus an age of 58 would be at the **64th percentile**.

b. Counting from the bottom of the list, there are 3 data values less than 25; and there is one data value equal to 25. So $x = 3$ and $y = 1$.

$$\Rightarrow \frac{x + 0.5y}{n} \cdot 100 = \frac{3 + 0.5(1)}{29} \cdot 100 = 12.07$$

Thus an age of 25 would be at the **12th percentile**.

z-Scores

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading. But we can tell which data value is farther removed from the center by finding how many standard deviations away from its mean the value is. And we have a statistic called a "*z*-score" for this purpose. Let x be a data value, μ be the mean, and σ the standard deviation. Then the z-score corresponding to x is:

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

Solving for x in this equation, we get $x = \mu + z\sigma$. So, this really does measure how many standard deviations x is from the mean. Moreover, if $z > 0$, a positive value, then x is greater than the mean; and if $z < 0$, a negative value, then x is less than the mean. When $z = 0$, the standard deviation is zero and the value is at the mean.

Example 3.17

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school.

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Which student had the highest GPA when compared to his school?

Solution to 3.17:

For each student, we use the z-score to determine how many standard deviations his GPA is away from the average for his school:

$$\text{For John, } z = \frac{x - \mu}{\sigma} = \frac{2.85 - 3.0}{0.7} = -0.21.$$

$$\text{For Ali, } z = \frac{x - \mu}{\sigma} = \frac{77 - 80}{10} = -0.3.$$

John's GPA is 0.21 standard deviations *below* his school's mean, while Ali's GPA is 0.3 standard deviations *below* his school's mean. So, John's *z*-score is higher than Ali's. For GPA, higher values are better; so, John has the better GPA relative to his school.

Try It Σ

3.14 Two swimmers, Angie and Beth, from different teams, wanted to find out who had the fastest time for the 50 meter freestyle when compared to her team. The table below shows their times.

Swimmer	Time in Seconds	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

Which swimmer had the fastest time when compared to her team?

Z- scores will be revisited in Chapter 7: The Normal Distribution. The normal distribution is an important continuous probability distribution. Z-scores will be on the x-axis and will also be used in hypothesis testing.

The figure below shows how z-scores relate to position to the mean and to percentile. Recall that the z-score at the mean is zero. The figure shows the normal distribution for the SAT. The z-score for Student A was 1, meaning Student A was 1 standard deviation above the mean. Thus, Student A performed in the 84.13 percentile on the SAT.

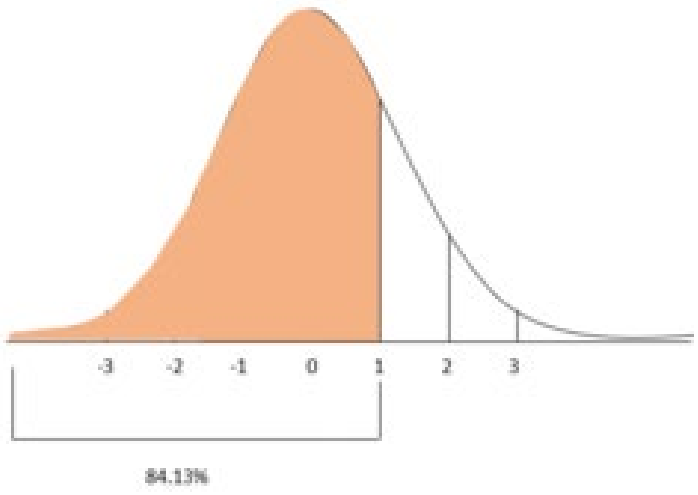


Figure 4.2 (credit: BaumannAshley, Wikimedia Commons)

Example 3.18

Three statistics students, Mia, Fiona and Raina, are in different classes and all scored an 82 on their last statistics test. In Mia's class the mean was 68 with a standard deviation of 6. In Fiona's class the mean was 73 with a standard deviation of 5. In Raina's class the mean was 80 with a standard deviation of 4. Who performed the best within their own class?

Solution 3.18

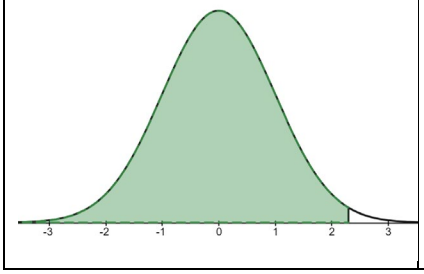
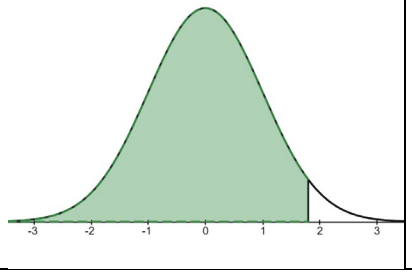
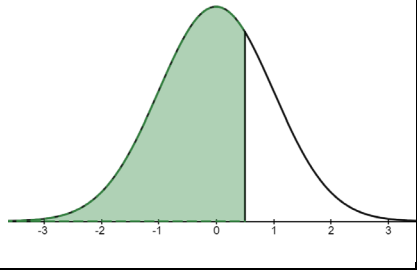
Even though they have the same score we do not know how they compare to their classmates so begin by calculating the z-scores for each student.

For Mia: $z = \frac{x-\mu}{\sigma} = \frac{82-68}{6} = 2.33$

For Fiona: $z = \frac{x-\mu}{\sigma} = \frac{82-73}{5} = 1.8$

For Raina: $z = \frac{x-\mu}{\sigma} = \frac{82-80}{4} = 0.5$

Comparing the graphs of the z-scores, it is easier to see that Mia scored the best within her class. The z-score for her score is at the highest percentile with the most area shaded to the left.

Mia: z-score 2.33	Fiona: z-score 1.8	Raina: z-score 0.5
		

KEY TERMS

Box plot a graph that gives a quick picture of the middle 50% of the data

First Quartile the value that is the median of the of the lower half of the ordered data set

Interquartile Range or *IQR*, is the range of the middle 50 percent of the data values; the *IQR* is found by subtracting the first quartile from the third quartile.

Mean a number that measures the central tendency of the data; a common name for mean is 'average.'
The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample is

$$\bar{x} = \text{sample mean} = \frac{\text{sum of data}}{\text{number of data values}} = \frac{\sum x}{n}$$

When the data is from a population, we denote the mean by the Greek letter μ.

Median the number that separates ordered data into halves; half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

Midpoint the number that is halfway between the endpoints of an interval

Mode the value that appears most frequently in a set of data

Outlier an observation that falls far outside the rest of the data.

Percentile a number that divides ordered data into hundredths; percentiles may or may not be part of the data. The median of the data is the second quartile and the 50th percentile. The first and third quartiles are the 25th and the 75th percentiles, respectively.

Quartiles the numbers that separate the data into quarters; quartiles may or may not be part of the data. The second quartile is the median of the data.

Standard Deviation: A numerical measure of dispersion in a data set. Specifically, this is the square root of the variance and measures how far data values are from their mean; notation: *s* for sample standard deviation and σ for population standard deviation.

Variance: The mean of the squared deviations from the mean, or the square of the standard deviation; for a set of data, a deviation can be represented as $x - \bar{x}$ where x is a data value and $\bar{x}$ is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.

FORMULA REVIEW

Mean $$\text{mean} = \frac{\text{sum of data}}{\text{number of data values}} = \frac{\sum x}{n}$$

Formulas for the Standard Deviation:

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

Population Standard Deviation

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

Sample Standard Deviation

A Formula for Finding the k^{th} Percentile

Suppose that we want the k^{th} percentile. It may or may not be part of the data.

Order the data, and let i be the index of the k^{th} percentile; i.e. this is the position that the percentile appears in the list of data. Let n be the total number of data values.

Then we calculate $i = \frac{k}{100}(n+1)$

If i is a positive integer, then the k^{th} percentile is the data value in the i^{th} position in the ordered set of data.

If i is not a positive integer, then round i up and round i down to the nearest integers.

Average the two data values in these two positions in the ordered data set. This is easier to understand in an example.

A Formula for Finding the Percentile of a Value in a Data Set

Again, order the data from smallest to largest.

Let x = the number of data values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile.

Let y = the number of data values equal to the data value for which you want to find the percentile.

Let n = the total number of data values.

Exercises for Chapter 3

For problems #1 - 4, find the mean, median, standard deviation and five number summaries for the given data set.

1. The following data shows the mileage (in mpg) for a random sample of 15 new cars.

30.1	32.8	29.2	32.6	29.0
28.2	31.5	32.8	32.8	30.2
31.5	32.9	31.4	30.6	31.0

2. The following data shows the ages (in years) of a random sample of 22 Boeing 747 airplanes.

17	5	5	12	8	5	8	16	14	12	22
15	5	8	17	5	4	2	22	17	20	23

3. The following data shows the speeds (in mph) for a random sample of 20 cars driving on I-294 at 1 pm on a single day.

68	65	50	79	77	60	55	61	78	75
75	67	72	58	70	62	67	72	70	74

4. The following data shows the number of cardiograms done each day for a random sample of 20 days at an outpatient testing center.

25	31	20	32	13	14	43	2	57	23
36	32	33	32	44	32	52	44	51	45

5. Find the mean and standard deviation for the following frequency tables:

a.

Grade	Frequency
49.5 – 59.5	2
59.5 – 69.5	3
69.5 – 79.5	8
79.5 – 89.5	12
89.5 – 99.5	5

b.

Daily Low Temperature	Frequency
49.5 - 59.5	53
59.5 - 69.5	52
69.5 - 79.5	15
79.5 - 89.5	1
89.5 - 99.5	0

c.

Points per Game	Frequency
49.5 - 59.5	14
59.5 - 69.5	32
69.5 - 79.5	15
79.5 - 89.5	23
89.5 - 99.5	2

6. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Calculate the following:

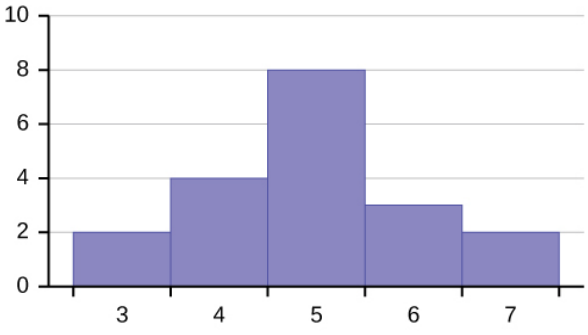
a. sample mean = _____ b) median = _____ c. mode = _____

7. When the data are skewed left, what is the typical relationship between the mean and median?

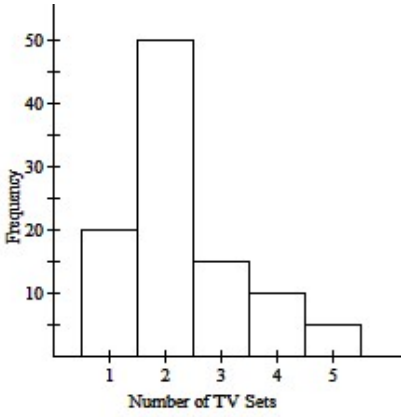
8. When the data are symmetrical, what is the typical relationship between the mean and median?

9. For each of the following histograms,

a) Describe the shape of the distribution
b) Describe the relationship between the mean and median.



(i)



(ii)

10. Which is the greatest, the mean, the mode, or the median of the data set?

11; 11; 12; 12; 12; 12; 13; 15; 17; 22; 22; 22

11. Which is the least, the mean, the mode, and the median of the data set?

56; 56; 56; 58; 59; 60; 62; 64; 64; 65; 67

12. Of the three measures, which tends to reflect skewing the most, the mean, the mode, or the median? Why?

13. In a perfectly symmetrical distribution, when would the mode be different from the mean and median?

14. The following data show the distances in miles between 20 retail stores and a large distribution center:

29; 37; 38; 40; 58; 67; 68; 69; 76; 86; 87; 95; 96; 96; 99; 106; 112; 127; 145; 150

a) Use a calculator to find the standard deviation and round to the nearest tenth.
b) Find the value that is one standard deviation below the mean.

For the next four exercises, calculate the mean, median and standard deviation:

15. The miles per gallon rating for 30 cars are shown:
19, 19, 19, 20, 21, 21, 25, 25, 25, 26, 26, 28, 29, 31, 31,
32, 32, 33, 34, 35, 36, 37, 37, 38, 38, 38, 38, 41, 43, 43

16. The height in feet of 25 trees is shown below (lowest to highest).

25, 27, 33, 34, 34, 34, 35, 37, 37, 38, 39, 39, 39,
40, 41, 45, 46, 47, 49, 50, 50, 53, 53, 54, 54

17. The following data are the prices of different laptops at an electronics store.

249, 249, 260, 265, 265, 280, 299, 299, 309, 319, 325, 326, 350,
350, 350, 365, 369, 389, 409, 459, 489, 559, 569, 570, 610

18. The following data are the daily high temperatures in a town for one month.

61, 61, 62, 64, 66, 67, 67, 67, 68, 69, 70, 70, 70, 71, 71,
72, 74, 74, 74, 75, 75, 75, 76, 76, 77, 78, 78, 79, 79, 95

19. Fredo and Karl, two baseball players on different teams, wanted to find out who had the higher batting average when compared to his team.

Player	Batting Average	Team Batting Average	Team Standard Deviation
Fredo	0.158	0.166	0.012
Karl	0.177	0.189	0.015

Which baseball player had the higher batting average when compared to his team?

20. The ages for 29 Academy Award winning best actors are shown below, in order from smallest to largest.

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47;
52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the 40th percentile.
b. Find the 78th percentile.
c. Find the percentile for age 37.
d. Find the percentile for age 72.

21. Jesse was ranked 37th in his graduating class of 180 students. At what percentile is Jesse's ranking?

22. On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

23. Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.

24. In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78th percentile. Should Li be pleased or upset by this result? Explain.

25. In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had $1,700 in damage and was in the 90th percentile. Should the manufacturer and the consumer be pleased or upset by this result? Explain and write a sentence that interprets the 90th percentile in the context of this problem.

26. Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is $240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?

27. The following data show the lengths of boats moored in a marina. Use this data to construct a boxplot.

16; 17; 19; 20; 20; 21; 23; 24; 25; 25; 25; 26; 26; 27;

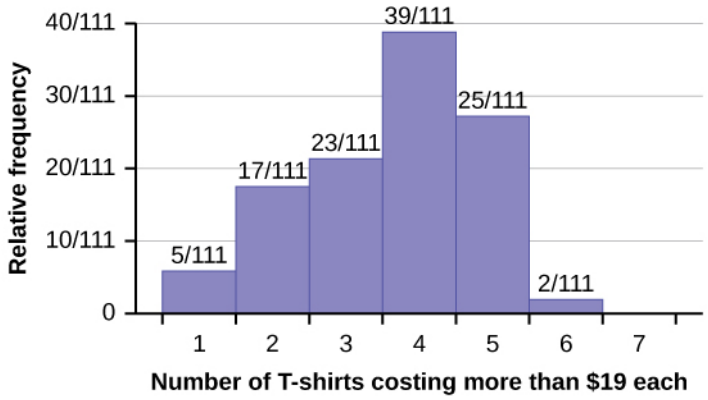
27; 27; 28; 29; 30; 32; 33; 33; 34; 35; 37; 39; 40

28. Find the five number summary and construct a boxplot for the given data set.

Amount (in dollars) spent on books for a semester

508	488	430	49	278	542	451	366	451	531	196	115
118	362	434	399	148	30	287	236	185	210	178	105
533	236	477	329								

Use the following information to answer the next two exercises: Suppose one hundred eleven people who shopped in a special t-shirt store were asked the number of t-shirts they own costing more than $19 each. The results are summarized in the histogram:



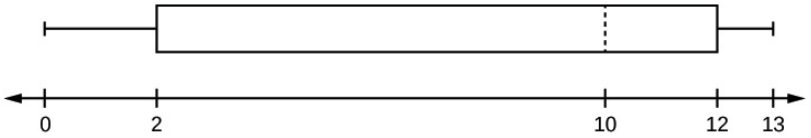
29. The percentage of people who own at most three t-shirts costing more than $19 each is approximately:

a. 21% b. 59% c. 41% d. Cannot be determined

30. If the data were collected by asking the first 111 people who entered the store, then the type of sampling is:

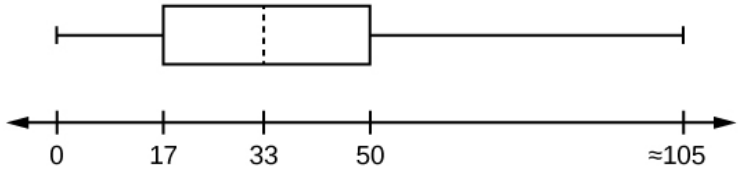
a. cluster b. simple random c. stratified d. convenience

31. Given the following box plot:



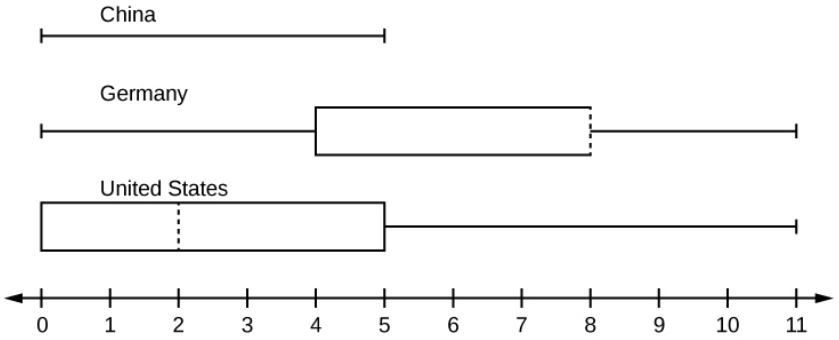
a. Which quarter has the smallest spread of data? What is that spread?
b. Which quarter has the largest spread of data? What is that spread?
c. Find the interquartile range (*IQR*).
d. Are there more data in the interval [5, 10] or in the interval [10, 13]? Explain

32. The following box plot shows ages of the U.S. population for 1990, the latest available year.



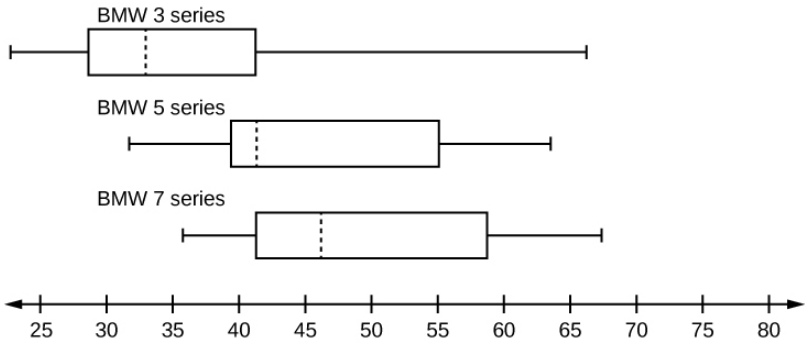
a. Are there fewer or more children (age 17 & under) than senior citizens (age 65 & over)? Explain.
b. Given that about 12.6% are age 65 and over, approximately what percentage of the population are working age adults (above age17 to age 65)?

33. In a survey of 20-year-olds in China, Germany, and the United States, people were asked the number of foreign countries they had visited in their lifetime. The following box plots display the results.



a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected.
b. Of the U.S. and Germany, which country has the greater percentage of 20-year olds that have visited more than eight foreign countries?
c. Compare the three box plots. What do they imply about the foreign travel of 20-year-old residents of the three countries when compared to each other?

34. A survey was conducted of 130 purchasers of new BMW 3 series cars, 130 purchasers of new BMW 5 series cars, and 130 purchasers of new BMW 7 series cars. In the survey, people were asked the age they were when they purchased their car. The following box plots display the results.



a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected for that car series.

b. Which group is most likely to have an outlier? Explain how you determined that.

c. Compare the three box plots. What do they imply about the age of purchasing a BMW from the series when compared to each other?

d. For the BMW 5 series, which quarter has the smallest spread of data? What is the spread?

e. For the BMW 5 series, which quarter has the largest spread of data? What is the spread?

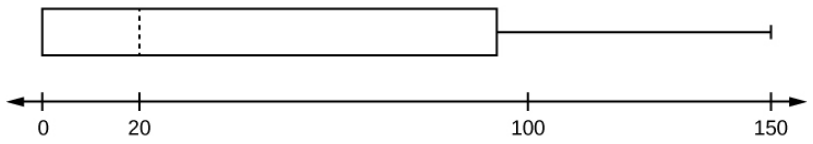
f. Estimate the interquartile range (IQR) for the BMW 5 series.

g. For the BMW 5 series, are there more data in the interval 31 to 38 or in the interval 45 to 55? How do you know this?

h. For the BMW 5 series, which interval has the fewest data in it? How do you know this?

i. 31 to 35 ii. 38 to 41 iii. 41 to 64

35. Given the following box plot:



What does it mean to have the first and second quartiles so close together, while the second to third quartiles are far apart?

36. The most obese countries in the world have obesity rates that range from 11.5% to 83.4%. This data is summarized in the following table:

Percent of Population Obese	Number of Countries
11.45–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

a. What is the best estimate of the average obesity percentage for these countries?
b. The United States has an average obesity rate of 33.9%. Is this rate above average or below?
c. What is the standard deviation for the listed obesity rates?
d. How does the United States compare to other countries? Would it be considered "unusual"?

37. The following table gives the percent of children under five considered to be underweight.

Percent of Underweight Children	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

a. What is the best estimate for the mean percentage of underweight children?
b. What is the standard deviation?
c. Which interval(s) could be considered unusual? Explain.

38. A music school has budgeted to purchase three musical instruments. They plan to purchase a piano costing $3,000, a guitar costing $550, and a drum set costing $600. The mean cost for a piano is $4,000 with a standard deviation of $2,500. The mean cost for a guitar is $500 with a standard deviation of $200. The mean cost for drums is $700 with a standard deviation of $100. Which cost is the lowest, when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type? Justify your answer.

39. Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best GPA when compared to other students at his school? Explain how you determined your answer.

Student	GPA	School Average GPA	School Std. Deviation
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

40. An elementary school class ran one mile with a mean of 11 minutes and a standard deviation of three minutes. Rachel, a student in the class, ran one mile in eight minutes. A junior high school class ran one mile with a mean of nine minutes and a standard deviation of two minutes. Kenji, a student in the class, ran 1 mile in 8.5 minutes. A high school class ran one mile with a mean of seven minutes and a standard deviation of four minutes. Nedda, a student in the class, ran one mile in eight minutes.

a. Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than he?
b. Who is the fastest runner with respect to his or her class? Explain.

41. The median age of the U.S. population in 1980 was 30.0 years. In 1991, the median age was 33.1 years.

a. What does it mean for the median age to rise?
b. Give two reasons why the median age could rise.
c. For the median age to rise, is the actual number of children less in 1991 than it was in 1980? Why or why not?

42. We are interested in the number of years students in a particular elementary statistics class have lived in California. The information in the following table is from the entire section.

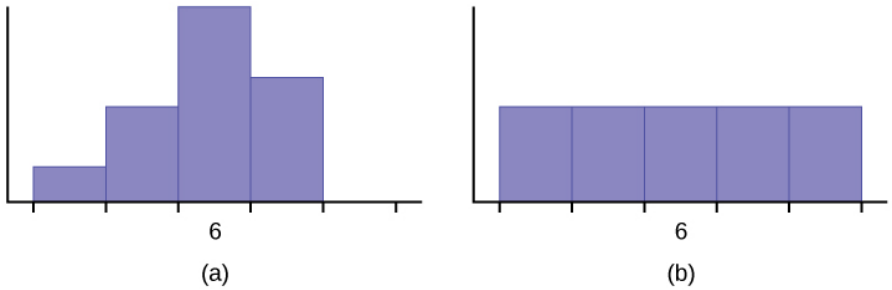
Number of years	Frequency	Number of years	Frequency
7	1	22	1
14	3	23	1
15	1	26	1
18	1	40	2
19	4	42	2
20	3		
			Total = 20

a. What is the *IQR*?
b. What is the mode?

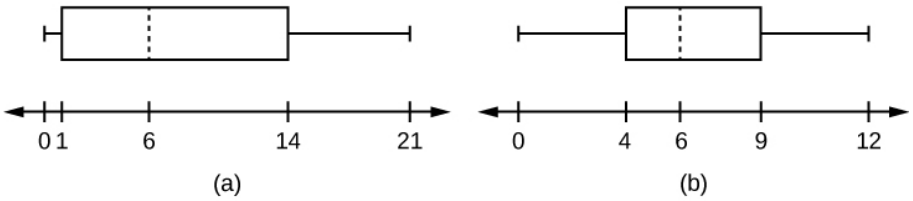
43. Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information:

	Javier	**Ercilia**
Sample mean $\bar{x}$	6.0 miles	6.0 miles
Sample SD, s	4.0 miles	7.0 miles

a. How can you determine which survey was more accurate?
b. Explain what the difference in the results of the surveys implies about the data.
c. If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?



d. If the two box plots depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?



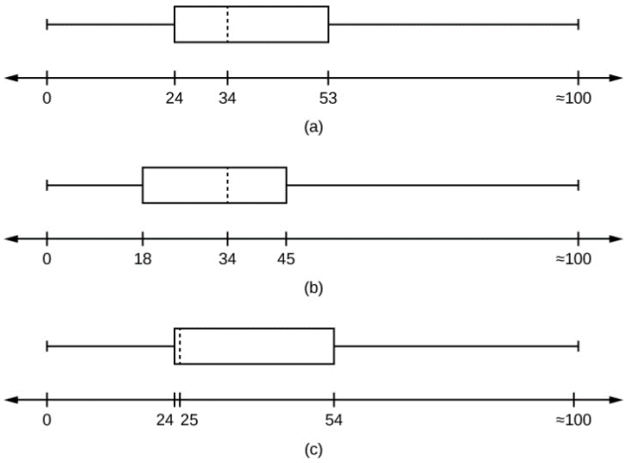
44. Santa Clara County, CA, has approximately 27,873 Japanese Americans. Their ages are as follows:

Age Group	**Percent of Community**
0 - 17	18.9
18 - 24	8.0
25 - 34	22.8
35 - 44	15.0
45 - 54	13.1
55 - 64	11.9
65+	10.3

a. Construct a histogram of the Japanese American community in Santa Clara County, CA. The bars will **not** be the same width for this example. Why not? What impact does this have on the reliability of the graph?

b. What percentage of the community is under age 35?

c. Which box plot most resembles the information above?

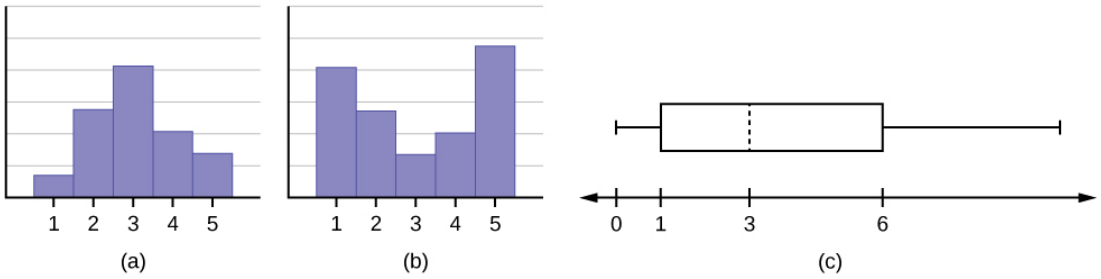


45. Forty randomly selected students were asked the number of pairs of sneakers they owned. Let X = the number of pairs of sneakers owned. The results are as follows:

X = # pairs of sneakers	Frequency	Relative Frequency
1	2	
2	5	
3	8	
4	12	
5	12	
6	0	
7	1	

a. Find the sample mean $\bar{x}$
b. Find the sample standard deviation, s.
c. Construct a histogram of the data.
d. Complete the third column of the chart.
e. Find the first quartile.
f. Find the median.
g. Find the third quartile.
h. Construct a boxplot for the data.
i. What percent of the students owned at least five pairs?
j. Find the 40th percentile.
k. Find the 90th percentile.

46. Refer to the graphs below determine which of the following are true and which are false. Explain your solution to each part in complete sentences.



a. The medians for all three graphs are the same.
b. We cannot determine if any of the means for the three graphs is different.
c. The standard deviation for graph b is larger than the standard deviation for graph a.
d. We cannot determine if any of the third quartiles for the three graphs is different.

47. Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a given year.

177; 205; 210; 210; 232; 205; 185; 185; 178; 210; 206; 212;
184; 174; 185; 242; 188; 212; 215; 247; 241; 223; 220; 260;
245; 259; 278; 270; 280; 295; 275; 285; 290; 272; 273; 280;
285; 286; 200; 215; 185; 230; 250; 241; 190; 260; 250; 302;
265; 290; 276; 228; 265

a. Find the five-number summary for the data.
b. Construct a box plot of the data.
c. The middle 50% of the weights are from ___ to ____.
d. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
e. If our population included every team member who ever played for the San Francisco 49ers, would the above data be a sample of weights or the population of weights? Why?
f. Assume the population was this particular year's San Francisco 49ers. Find:
   i. the population mean, μ.
   ii. the population standard deviation, σ.
   iii. the weight that is two standard deviations below the mean.
   iv. When quarterback Steve Young played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?

48. In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

a. Organize the data in a relative frequency table.
b. Find the median, the first quartile, and the third quartile.
c. Find the 65th percentile.
d. The middle 50% of the conferences last from _days to _____days.
e. Construct a box plot of the data.
f. Find the 10th percentile.
g. Calculate the sample mean of days of engineering conferences.
h. Calculate the sample standard deviation of days of engineering conferences.
i. Find the mode.
j. If you were planning an engineering conference, which would you choose as the length of the conference: mean, median, or mode? Explain why you made that choice.
k. Give two reasons why you think that three to five days seem to be popular lengths of engineering conferences.

49. A survey of enrollment at 35 community colleges across the United States yielded the following data:

6414;	1550;	2109;	9350;	21,828;	4300;	5944;	5722;	2825;
2044;	5481;	5200;	5853;	2750;	10,012;	6357;	27,000;	9414;
7681;	3200;	17,500;	9200;	7380;	18,314;	6557;	13,713;	17,768;
7493;	2771;	2861;	1263;	7285;	28,165;	5080;	11,622	

a. Make a frequency table for the data using five intervals of equal width.
b. Construct a histogram of the data.
c. If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
d. Calculate the sample mean.
e. Calculate the sample standard deviation.
f. A school with an enrollment of 8000 would be how many standard deviations away from the mean?

50. The average height of US men in 2010 is 69.3 inches according to Medical News Today. Let's say the standard deviation is 2.8 inches.

a. Using Chebyshev's theorem, at least 75% of heights will be between what two heights?
b. Using Empirical rule, 68% of heights will be between what two heights?
c. If a randomly chosen male has a height of 78 inches, what is his z-score?
d. Using Empirical rule, what percentage of heights are above 74.9 inches?

51. The average ACT score in 2010 was 21 with a standard deviation of 5.2 according to Digest of Education Statistics.

a. Using Chebyshev's theorem, at least 89% of scores will be between what two scores?
b. Using Empirical rule, 95% of scores will be between what two scores?

c. If a randomly chosen student has a score of 18, what is this student's z-score?
d. Using Empirical rule, what percentage of scores are below 26.2?

52. Let's say that the mean return of mutual funds in the last quarter is 2.8% with a standard deviation of 4.5%. Assume the returns are symmetrically distributed. Find the return value(s) that would separate the

a. top 2.5%
b. middle 99.7%
c. bottom 84%

53. Let's say that the mean return of mutual funds in the last quarter is 2.8% with a standard deviation of 4.5%. Assume the returns are symmetrically distributed.

a. Find the percent of returns that are above 7.3%.

b. Find the percent of returns below 11.8%.

54. Let's say that the mean return of mutual funds in the last quarter is 5.8% with a standard deviation of 2.1%. Assume the returns are skewed right.

a. Find the percent of returns that are between 1.6% and 10%.

b. At least 89% of returns are between what two values?

REFERENCES

3.1 Measures of the Center of the Data

Data from The World Bank, available online at http://www.worldbank.org (accessed April 3, 2013).

"Demographics: Obesity – adult prevalence rate." Indexmundi. Available online at http://www.indexmundi.com/g/
r.aspx?t=50&v=2228&l=en (accessed April 3, 2013).

3.2 Measures of Variation

Data from Microsoft Bookshelf.

King, Bill."Graphically Speaking." Institutional Research, Lake Tahoe Community College. Available online at http://www.ltcc.edu/web/about/institutional-research (accessed April 3, 2013).

Villines, Z. "What is the average height for men?" Medical News Today. Available online at https://www.medicalnewstoday.com/articles/318155.php (accessed May 20, 2018).

ACT Scores data from https://nces.ed.gov/programs/digest/d10/tables/dt10_155.asp

3.3 Measures of Relative Position

Cauchon, Dennis, Paul Overberg. "Census data shows minorities now a majority of U.S. births." USA Today, 2012. Available online at http://usatoday30.usatoday.com/news/nation/story/2012-05-17/minority-birthscensus/55029100/1 (accessed April 3, 2013).

Data from the United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/ (accessed April 3, 2013).

"1990 Census." United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/main/www/cen1990.html (accessed April 3, 2013).

Data from *San Jose Mercury News*.

Data from *Time Magazine*; survey by Yankelovich Partners, Inc.

Data from *West Magazine*.

Introductory Statistics, Sawyer Academy, 2012 (https://saylordotorg.github.io/text_introductory-statitstics/s06-05-the-empirical-rule-and-chebysh.html) accessed on 7/15/2021.

CHAPTER 3 SOLUTIONS:

1) $\bar{x}$ = 31.1, med = 31.4, s = 1.54, min = 28.2, Q1 = 30.1, Q2 = 31.4, Q3 = 32.8, max = 32.9

3) x = 67.75, med = 69, s = 8.04, min = 50, Q1 = 61.5, Q2 = 69, Q3 = 74.5, max = 79

5) a. x bar = 79.5, s = 11.1 b. x bar = 61.5, s = 7.1 c. x bar = 70.7, s = 11.2

7) The mean is less than the median

9) a. symmetric b. skewed to the right

11) mode

13) If the distribution was bimodal, the mode would be different from the mean and median.

15) x bar = 30.7, med = 31.5, s = 7.5

17) x bar = 371.3, med = 350, s = 109.9

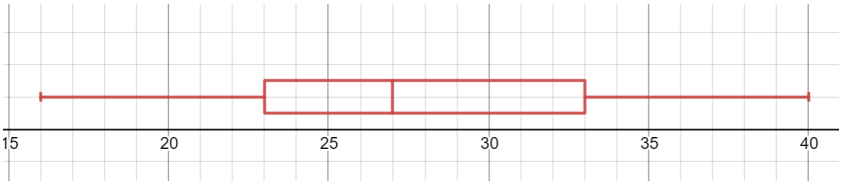
19) Fredo, Fredo's z score = -67 (less negative), Karl's z score = -.8

21) 94th percentile

23) Being at the 85th percentile of wait times is bad as 85 percent of the people have been waiting less time than Mina. Only 15 percent will be waiting for longer.

25) $1,700 is in the 90% percentile of repair costs so it is comparatively high. The manufacturer and the consumer would not be pleased. Only 10% of cars in the crash test had higher repair costs.

27) min = 16, Q1 = 23, med = 27, Q3 = 33, max = 40



29) c

31) a. Q4 has a spread of 1 b. Q2 has a spread of 8 c. 12 –2 = 10 d. There is more data in [10,13] as it contains half of the data, both Q3 and Q4. [5,10] is within Q2 of the data.

33) a. From the survey of 20-year-olds from China, all the surveyed people have been between 0 and 5 countries. The 20-year-olds from Germany data is symmetric where the

middle 50% of the data is between 4 and 8 countries. The 20-year-olds from the US are right skewed with the middle 50% of the data between 0 – 5 countries.

b. Germany, the 4^{th} quartile is 8 countries and above.

c. The survey implies that 20-year-olds from Germany have visited more countries compared to the US and China. It also implies that 20-year-olds from China have visited fewer countries comparatively.

35) Since the first and second quartile, so close together a quarter of the data falls within that interval. Since the second and third are far apart, the data is more spread out in that quartile.

37) a. estimated mean = 26.573 b. s = 8.540 c. The last interval could be considered unusual because it is greater than 2 standard deviations above the mean (26.573 + 2[8.540] = 43.653)

39) $Z_{Thuy} = -.625$, $Z_{Vichet} = 0.60$, $Z_{Kamala} = 1.5$, Kamala has the highest GPA compared to other students in her school.

41) a. When the median age increases the middle age of the US is older. More of the population is older.

b. Two possible reasons are that the birth rate is not increasing as fast as the death rate or that people are living longer.

c. Not necessarily, it does not have to be less children in 1991. It just has to 50% of the people older than 33.1 and 50% below. It could be that people were living longer.

43) a. Since there are only two samples taken with the same n value, it would be important to have information about sampling and take more samples. Taking more samples would help determine which sample is more representative of the population.

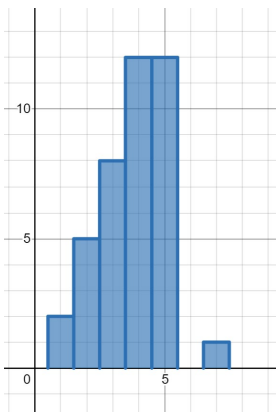
b. The mean is the same, but Ercilia's data implies customers live farther from the mall.

c. (b) since the data has a greater range, and Ercilia has the greater standard deviation.

d. (a) since the data has a greater range, and Ercilia has the greater standard deviation.

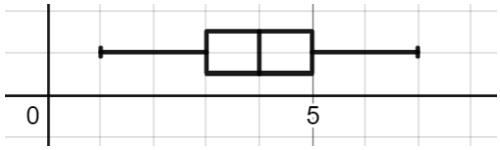
45) a. x bar = 3.8 b. s = 1.3

c.



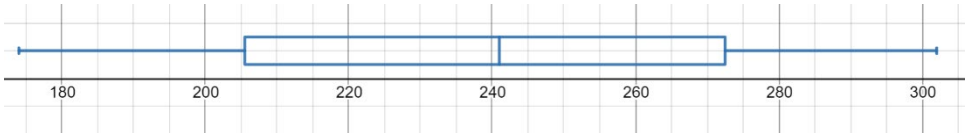
d. 2/40 (0.05 or 5%), 5/40 (0.125 or 12.5%), 8/40 (0.2 or 20%), 12/40 (0.3 or 30%), 12/40 (0.3 or 30%), 0/40 (0 or 0%), 1/40 (0.025 or 2.5%)

e. Q1 = 3 f. med = 4 g. Q3 = 5

h. 

I. 13/40 = 0.325 x 100 = 32.5% j. 40th percentile = 4 k. 90th percentile = 5

47) a. min = 174, Q1 = 205.5, med = 241, Q3 = 272.5, max = 302

b. 

c. 205.5 to 272.5

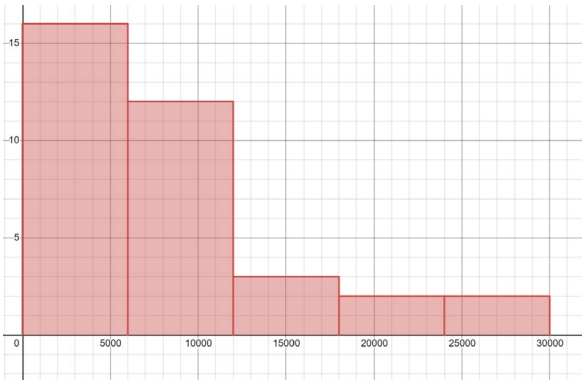
d. sample of weights; it's just one team

e. sample of weights; there are only 53 data points given, and there have been more than 53 San Francisco 49ers players since the team began

f. (no d) i. 236.3 ii. 37.5 iii. 161.3 iv. Zsy = -.81, 0.81 standard deviations below the mean.

49) a.

Interval	Frequency
1 - 6000	16
6001 - 12000	12
12001 - 18000	3
18001 - 24000	2
24001 - 30000	2

b. 

c. the mode d. mean = 8631.6 e. 6944.1 f. Z = -0.091, 0.091 standard deviations below the mean

51) a. between 5.4 and 36.6 b. between 10.6 and 31.5 c. Z = -0.58 d. 84%

53) a. 16% b. 97.5%