# 3 – 1 Measures of Central Tendency

## Objective 1. Summarize Data, Using Measures of Central Tendency, Such As the Mean, Median, Mode, and Midrange.

Measures of Central Tendency are also called averages. Notice the purpose of each and how to use them to describe the data.

### Parameters and Statistics

Recall the definitions of a population and sample:

***Population*** – all subjects of interest or being studied. Subjects may be humans, animals, business entities, products, etc. The size of the population usually makes gathering data from every subject difficult due to expense or time.

***Sample*** – a group of *some* subjects selected from the population.

Measures found by using all data values in a population are called \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

Measures found by using data values from a sample are called \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

### General Rounding Rule

In statistics, round only after the \_\_\_\_\_\_\_\_\_\_ answer is calculated. Intermediate rounding increases the difference between the answer and the exact value of the statistic.

## The Mean

The **mean** is the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_, found by adding the values of the data and dividing by the total number of values.

Round the calculated value of the mean to one more decimal place than occurs in the raw data.

The \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ mean, denoted by \_\_\_\_\_ (pronounced “mew”), is calculated by using all values in the population. The population mean is a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

where *X* represents a data value in the population and N represents the total number of values in the population.

The \_\_\_\_\_\_\_\_\_\_\_\_\_\_ mean, denoted by \_\_\_\_\_ (pronounced “*x*-bar”), is calculated by using the data values in the sample. The sample mean is a \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

where *X* represents a data value in the sample and *n* represents the total number of values in the sample.

In statistics, Greek letters denote \_\_\_\_\_\_\_\_\_\_\_\_\_\_ and Roman letters denote \_\_\_\_\_\_\_\_\_\_\_\_\_\_. Assume that the data are obtained from samples unless otherwise specified.

### Example 3-1. Salaries of College and University Coaches

The data are salaries (in hundred thousands of dollars) of a sample of 30 college and university coaches in the United States. Find the sample mean. (Use the technology of your choice.)

164 225 225 140 188

210 238 146 201 544

550 188 415 261 164

478 684 330 307 435

857 183 381 275 578

450 385 297 390 515

*Solution:*

The sample mean salary of college and university coaches is \_\_\_\_\_\_\_\_\_\_\_\_.

In the previous chapter we constructed a frequency chart with 8 classes for this sample. Use the frequency chart to calculate the mean for grouped data. This process is most frequently used when the raw data is not available.

, where f is the frequency and is the midpoint of the class.

| **Class Limit** | **Class Boundary** | **Class Midpoint** | **Frequency** | **f\*Xm** |
| --- | --- | --- | --- | --- |
| 140 - 229 | 139.5-229.5 | 184.5 | 11 |  |
| 230 – 319 | 229.5 – 319.5 | 274.5 | 5 |  |
| 320 – 409 | 319.5 – 409.5 | 364.5 | 4 |  |
| 410 – 499 | 409.5 – 499.5 | 454.5 | 4 |  |
| 500 – 589 | 499.5 – 589.5 | 544.5 | 4 |  |
| 590 – 679 | 589.5 – 679.5 | 634.5 | 0 |  |
| 680 - 769 | 679.5 – 769.5 | 724.5 | 1 |  |
| 770 - 859 | 769.5 – 859.5 | 814.5 | 1 |  |
| **Totals** | | | **30** |  |

Find the sample grouped mean for the frequency chart. \_\_\_\_\_\_\_\_

Compare the sample means calculated from the raw sample data and from the grouped data in the frequency chart.

**Note:** The mean (is always / is not always) an actual data value.

### Review of Properties of the Mean

1. The mean is found by using \_\_\_\_\_\_\_\_ values of the data set.
2. The mean of a data set is \_\_\_\_\_\_\_\_\_\_\_\_\_\_.
3. The mean of a data set is not necessarily one of the \_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_.
4. To find the mean of a frequency distribution, there must not be an \_\_\_\_\_\_\_\_\_\_\_\_\_ class.
5. The mean is affected by extremely high or low values, called \_\_\_\_\_\_\_\_\_\_\_.

### The Median

The **median** is the \_\_\_\_\_\_\_\_\_\_\_ of an ordered list of the data (the data array). The symbol for median is MD, or Med.

To find the median,

1. Arrange the data from smallest to largest;
2. Determine the number of values, *n*, in the data set;
3. a. If *n* is odd, the middle data value is the median; or

b. If *n* is even, the median is the mean of the middle two values.

### Example 3.2. Calories per Ounce of Salad Dressing

In the previous chapter, we ordered the data for a sample of calories per 1 ounce of 36 salad dressings (not fat-free).

100 100 100 100 100 110 110 115 120 120 120 120 120 120 130 130 130 130 130 130 140 140 140 140 145 145 145 150 150 160 160 160 160 160 160 170

*Solution:*

Find the median of this data. Med = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

### Example 3-3. Number of Tornados per Year

NOAA (National Oceanic and Atmospheric Administration: National Centers for Environmental Information) reports the number of tornados per year.

2192 1156 1282 1625 878 891 881

*Solution:*

Find the median. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

### Review of Properties of the Median

1. The median is used to find the \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ of a data set.
2. The median is used when one needs to determine if a data value falls in the \_\_\_\_\_\_\_\_ half or \_\_\_\_\_\_\_\_\_ half the distribution.
3. The median is not affected very much by \_\_\_\_\_\_\_\_\_\_\_\_\_.
4. The median can be used with \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ distributions.

### The Mode

The **mode** is the value in the data set that occurs \_\_\_\_\_\_\_\_\_\_\_ often.

Data sets that have only one value that has the \_\_\_\_\_\_\_\_\_\_ frequency are called **unimodal.**

### Example 3-4. Find the Mode of a Unimodal Sample

Find the mode: 18.0, 14.0, 34.5, 10.0, 11.3, 10.0, 12.4, 10.0

*Solution:*

There is \_\_\_\_\_\_\_\_\_ mode, so the sample is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

\_\_\_\_\_\_\_\_ occurs \_\_\_\_\_\_\_ times, which is more than any other data value.

Data sets that have exactly two values that occur with the same \_\_\_\_\_\_\_\_\_\_\_\_ frequency are said to be **bimodal.**

### Example 3-5. Find the Modes of a Bimodal Sample

Find the two modes in the following bimodal data set.

104 107 109 104 109 111 104 109

112 111 104 109 104 110 109

*Solution:*

When a data set has more than two values that occur with the same \_\_\_\_\_\_\_\_\_\_\_\_ frequency, it is called **multimodal.**

### Example 3-6. Find the Modes of a Multimodal Sample

In **Example 3.2,** three values occur 6 times, which is the greatest frequency, which means this data set is \_\_\_\_\_\_\_\_\_\_\_\_\_.

100 100 100 100 100 110 110 115 120 120 120 120 120 120 130 130 130 130 130 130 140 140 140 140 145 145 145 150 150 160 160 160 160 160 160 170

*Solution:*

There are \_\_\_\_\_\_\_\_\_ modes which are \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

Some data sets have **no mode** at all. This happens when all of the values occur at the same frequency.

The data in **Example 3.3** has **no mode**. All values in the sample occur exactly once. 2192 1156 1282 1625 878 891 881

The data set has no mode.

The mode is the only measure of central tendency that can be used to find the most typical case for nominal data.

### Example 3-7. Sales for Fast-Food Franchises

From Example 2-7, we have nominal data of the worldwide sales, in billions of dollars, for fast-food franchises for a specific year.

|  |  |
| --- | --- |
| Wendy’s | $ 8.7 |
| KFC | 14.2 |
| Pizza Hut | 9.3 |
| Burger King | 12.7 |
| Subway | 10.0 |
| Source: Franchise Times | |

*Solution:*

The category with the greatest frequency is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_. Thus, the most typical case is \_\_\_\_\_\_\_\_\_\_\_\_\_\_, meaning the mode is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

### Review of Properties of the Mode

1. The mode is used to find the most \_\_\_\_\_\_\_\_\_\_\_\_ case.
2. The mode is the \_\_\_\_\_\_\_\_\_\_\_\_\_\_ average to find.
3. The mode can be used for \_\_\_\_\_\_\_\_\_\_\_\_\_\_ and \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ variables.
4. The mode is not necessarily \_\_\_\_\_\_\_\_\_\_\_\_\_\_ and may not even \_\_\_\_\_\_\_\_\_\_\_.

### Resistant Measures of Central Tendency

A measure is resistant when an extremely high or extremely low data value does not change the value of that measure. The \_\_\_\_\_\_\_\_\_\_\_ and \_\_\_\_\_\_\_\_\_\_\_ are resistant. However, extreme values, also called outliers, can have a striking effect on the \_\_\_\_\_\_\_\_\_\_\_\_. (We will learn how to determine which values are outliers later in this chapter.)

### Example 3-8. Find the Measures of Central Tendency

The following data set has a value much higher than the rest:

177 153 122 141 189 155 162 165 149 157 240

When the large value is deleted from the data set, the data set becomes:

177 153 122 141 189 155 162 165 149 157

Find the measures of central tendency, then compare and contrast the values.

*Solution:*

The measures of central tendency for the first data set follow:

The mode is \_\_\_\_\_\_\_\_\_\_\_.

The median is \_\_\_\_\_\_\_\_\_\_.

The mean is \_\_\_\_\_\_\_\_\_\_.

The measures of central tendency for the data set without the higher value are:

Now the mode is \_\_\_\_\_\_\_, the median is \_\_\_\_\_\_\_\_\_ and the mean is \_\_\_\_\_\_\_\_\_.

Notice how much the mean changed (The mean is higher for the original data set and lower for the edited data set.) and how close the median of the edited data set is to the original value.

### The Midrange

The **midrange (MR)** is the arithmetic mean of the lowest and highest values in the data set. It is the midpoint of the range. It is not efficient, nor resistant. However, it can estimate the center for uniform distributions.

### Example 3-9. Finding the Midrange

Find the midrange of each of the data sets.

Original Data Set:

177 153 122 141 189 155 162 165 149 157 240

Edited Data Set:

177 153 122 141 189 155 162 165 149 157

*Solution:*

Original Data Set: MR = \_\_\_\_\_\_\_\_\_\_\_\_

Now find the midrange for the data set after deleting the extreme value.

Edited Data Set: MR = \_\_\_\_\_\_\_\_\_\_\_\_

Is it reasonable to use the midrange as a descriptive statistic for this data set?

### Review of Properties of the Midrange

1. The midrange is \_\_\_\_\_\_\_\_\_\_to compute.
2. The midrange gives the \_\_\_\_\_\_\_\_\_\_\_\_\_ of the range of the data set.
3. The midrange is affected by \_\_\_\_\_\_\_\_\_\_\_\_\_\_.

### Shapes of Distributions

In chapter 2, many different shapes of distributions were discussed. The three most important shapes to consider are those that are positively skewed, symmetric and negatively skewed.

For positively skewed or right-skewed distributions, most of the data falls to the \_\_\_\_\_\_\_ of the mean and clusters at the lower end of the distribution. The tail stretches out to the \_\_\_\_\_\_\_\_\_. The mean will be to the right of the median and the mode will be on the left. The mean is “pulled” toward the tail by the more extreme values.

The graph shows a histogram with the Mode less than both the Mean and Median; the Median less than the Mean, but more than the Mode; and the Mean greater than both the Mean and Median.

For negatively skewed or left-skewed distributions, most of the data falls to the \_\_\_\_\_\_\_ of the mean and clusters at the upper end of the distribution. The tail stretches out to the \_\_\_\_\_\_\_\_\_. The mean will be to the left of the median and the mode will be on the right. As before, the mean is “pulled” toward the tail by the more extreme values.

The graph shows a histogram with the Mode greater than both the Mean and Median; the Median greater than the Mean, but less than the Mode; and the Mean less than both the Mean and Median.

For symmetric distributions, the data values are distributed evenly on \_\_\_\_\_\_ \_\_\_\_\_\_\_ of the mean. The distribution is \_\_\_\_\_\_\_\_\_\_\_\_, with the mean, mode and median are equal and at the center of the distribution.

The graph shows a bell-shaped histogram with equal Mode, Mean and Median.

# 3 – 2 Measures of Variation

## Objective 2. Describe Data, Using Measures of Variation, Such As the Range, Variance, and Standard Deviation.

### Example 3-10. Fading Time of Outdoor Paint

Consider the sample of time, in months, two brands of outdoor paint last before fading.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Brand A** | 10 | 60 | 50 | 30 | 40 | 20 |
| **Brand B** | 35 | 45 | 30 | 35 | 40 | 25 |

Calculate the sample mean time, in months, to fade for each brand.

*Solution:*

Mean for Brand A: = \_\_\_\_\_ Mean for Brand B: = \_\_\_\_

Notice that we need more information is needed to describe the data.

### Range

The range is the simplest of the measures of variability to calculate. It is the difference between the highest value minus the lowest value.

### Example 3-11. Find the Range

Find the range for Brand A.

Find the range for Brand B.

*Solution:*

Find the range for Brand A: RA = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ = \_\_\_\_\_ months

Find the range for Brand B: RB = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ = \_\_\_\_\_ months

Although the means are equal for both brands, the ranges are quite different.

### Population Variance and Standard Deviation

Statisticians use the standard deviation as a more meaningful statistic than the range to measure variability.

Data variation is based on the difference between each data value and the mean. This difference is called a \_\_\_\_\_\_\_\_\_\_\_\_\_.

The deviation of 10 from the mean of the sample of time to fading of Brand A is \_\_\_\_\_\_\_\_\_= \_\_\_\_\_\_.

The sum of the deviations of **all** data values in the population about the population mean, without rounding, will be zero. To eliminate this problem, we square the deviations, add the squares and find the mean of the sum of the squares by the total number of data values. This is called the **population variance**, symbolized by . However, this measure is in the square of the units of measure of the data. Statisticians take the square root of the variance and call it the **population** **standard deviation**, .

, where

*X* = each individual data value in the population  
*N* = population size, and   
= population mean.

Thus, the population variance is .

Normally, these parameters are calculated using technology.

**Round the population standard deviation to at least one more decimal place than that of the original data.**

### Sample Variance and Standard Deviation

More often the sample standard deviation is used because sample data is available more often than data for an entire population.

The formula for the sample standard deviation is used to estimate the population standard deviation. However, merely using the sample mean, , instead of the population mean, , and the sample size, *n*, instead of the population size, *N*. This would underestimate the population standard deviation. Therefore, to calculate an unbiased estimate of the population variance or population standard deviation, divide by , instead of *n:*

, where

*X*= each individual data value in the sample  
*n* = sample size, and   
= sample mean.

Thus, the sample variance is .

### Example 3-12. Find the Sample Standard Deviation for Brand A

Find the sample standard deviation for Brand A, listed in Example 3.9.

Recall the sample mean for Brand A: = \_\_\_\_\_

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Brand A** | 10 | 60 | 50 | 30 | 40 | 20 |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

Solution, by hand:

or, using the shortcut formula that does not use the mean of the sample,

*Solution, using the TI84 calculator:*

Enter the six data values into one list.

<STAT> 1:Edit Select list and enter data.

<STAT> -> CALC 1: OneVarStats

\* Enter screen of entry

\* Enter screen of results

Now find the standard deviation of Brand B. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

What does this indicate about the variation of the two brands?

The sample standard deviation is Sx.  
If a population is entered, the population standard deviation is σx.

### Uses of the Standard Deviation

1. The standard deviation provides a measure of \_\_\_\_\_\_\_\_ of the data. If the standard deviation is large, the data are more \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_. It is used to compare the variability of two or more data sets.
2. The standard deviation is used to determine the consistency of a \_\_\_\_\_\_\_\_\_\_\_\_.
3. The standard deviation is used to find the number of data values that fall in a specific interval in a distribution.
4. The standard deviation is used in inferential statistics.

### Coefficient of Variance

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ is a statistic that allows comparison of standard deviations of data sets having different units, such as a variable measured in inches and a variable measured in centimeters.

The formula for the sample Coefficient of Variance, CVar, is expressed as a percentage:

### Range Rule of Thumb

A rough estimate of the standard deviation is

### Chebyshev’s Theorem

The standard deviation of a variable measures spread, or dispersion, of the variable. The larger the standard deviation, the greater the dispersion. The smaller the standard deviation, the less the data values are dispersed and the data values are more consistent. Chebyshev’s Theorem finds the proportion of spread in terms of standard deviations. *The theorem applies to any distribution regardless of shape.*

### Chebyshev’s Theorem

The proportion of values from a data set that fall within *k* standard deviations of the mean will be at least , where k >1 and not necessarily an integer.

### Example 3-13. Hours Online Per Day

Americans spend an average of 2.85 hours per day online. If the standard deviation is 32 minutes, find the range in which at least a) 75% (that is, when *k* = 2 standard deviations) and b) 88.89% (that is, when *k* = 3 standard deviations) of the data will lie.

(Source: [Internet Advertising Bureau, U.K.](https://www.iabuk.net/about/press/archive/definitive-time-people-spend-online-2hrs-51-mins-a-day), 2015.)

*Solution:*

First, change the average to the same units as the standard deviation.

2.85 hours = \_\_\_\_\_\_\_\_\_\_minutes (Multiply 2.85 hours by 60 minutes/hour)

a) For *k* = 2, =

171 + 2(32) = 171 + 64 = \_\_\_\_\_\_\_ minutes or \_\_\_\_\_\_\_\_ hours

171 – 2(32) = 171 – 64 = \_\_\_\_\_\_\_ minutes or \_\_\_\_\_\_\_\_\_ hours

Thus, by Chebyshev’s Theorem, at least 75% of the data is between \_\_\_\_\_\_\_\_ and \_\_\_\_\_\_\_\_\_ hours.

b) For *k* = 3, =

171 + 3(32) = 171 + 96 = \_\_\_\_\_\_ minutes or \_\_\_\_\_\_ hours

171 – 3(32) = 171 – 96 = \_\_\_\_\_\_ minutes or \_\_\_\_\_\_ hours

Thus, by Chebyshev’s Theorem, at least 88.89% of the data is between \_\_\_\_\_\_ and \_\_\_\_\_\_ hours.

### Example 3-14. Average Time to Traverse a Maze

The average number of trials it took a sample of mice to learn to traverse a maze was 12. The standard deviation was 3. Use Chebyshev’s Theorem to find the minimum percentage of data values that fall between 4 and 20 trials.

*Solution:*

12 - 3k = 4 Solve for *k* to find the number of standard deviations 4 is from the mean.

*k* = \_\_\_\_\_

12 + 3k = 20 Solve for k to find the number of standard deviations 20 is from the mean.

*k* = \_\_\_\_\_

You can use Chebyshev’s Theorem as 4 and 20 are the same number of standard deviations from the mean.

Use the formula to find the percentage: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

### The Empirical (Normal) Rule

While Chebyshev’s Theorem applies to any distribution, regardless of shape, the **Empirical Rule** applies to *bell-shaped* data.

By the Empirical Rule, for a distribution that is bell-shaped,

1. Approximately 68% of the data values will fall within 1 standard deviation of the mean.
2. Approximately 95% of the data values will fall within 2 standard deviation of the mean.
3. Approximately 99.7% of the data values will fall within 3 standard deviation of the mean.

A symmetric bell-shaped histogram having the mean in the middle of the distribution and the positions of 1, 2, and 3 standard deviations above and below the mean marked.  The amount of the distribuiton within 1 standard deviation of the mean is marked as about 68% of the data.  The amount of the distribuiton within 2 standard deviations of the mean is marked as about 95% of the data.  The amount of the distribuiton within 3 standard deviations of the mean is marked as about 99.7% of the data.

### Example 3-15. Average Math and Reading SAT Scores

The average Math and Reading SAT score in 2014 was 1010. Suppose the distribution of scores is approximately bell shaped and the standard deviation is approximately 86. a) Within what boundaries would you expect 68% of the scores? b) What percentage of scores would you expect to be above 1182?

*Solution:*

1. I would expect 68% of the data values to fall within 1 standard deviation of the mean.

The upper boundary will be \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The lower boundary will be \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Thus, I expect 68% of the data values to be between \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

1. To find the percentage of scores above *1182*, find the number of standard deviations the value is from the mean.

1182 = 1010 + (number of standard deviations from the mean) \* 86

172 = (number of standard deviations from the mean) \* 86

2 = number of standard deviations from the mean

By the Empirical Rule, we expect 95% of the data within 2 standard deviations of the mean. So, 5% is evenly split between the amount of data values above and below the mean. That means 2.5% of the data is expected to be above 1182.

# 3 – 3 Measures of Position

Objective 3. Identify the Position of a Data Value in a Data Set, Using Various Measures of Position, Such as Percentiles and Quartiles.

Measures of position are used to locate the \_\_\_\_\_\_\_\_\_\_\_\_ position of a data value in a data set. A value at the 88th percentile has \_\_\_\_\_\_% of the data values below it and \_\_\_\_\_\_% above it. The median is a measure of position corresponding to the 50th percentile.

### Standard Scores

Standard scores can be used to compare the position of data values on two different distributions by providing a measure of how many standard deviations a particular data value is from the mean.

for samples and , for populations where

*X* = sample data value and *X* = population data value;

= sample mean and population mean; and

sample standard deviation and population standard deviation.

### Example 3-16. Compare Two Distributions

Did a student perform better on a calculus test with a score of 65 having a mean of 50 and standard deviation of 10 or on a history test with a score of 30 having a mean of 25 and a standard deviation of 5?

*Solution 1*:

65 and 30 are both above the mean. 65 is 15 points above the mean of the calculus test, which is one and a half standard deviations above the mean. 30 is 5 points above the mean of the history test, which is one standard deviation above the mean. The student has a higher relative position on the \_\_\_\_\_\_\_\_ test.

*Solution 2*:

We can use the z score or standard score to determine how many standard deviations above or below the mean a data value is.

For the calculus test, meaning 1.5 standard deviations above the mean.

For the history test, meaning 1 standard deviation above the mean.

The z score for the calculus test is larger, so the relative position in the \_\_\_\_\_\_\_\_ class is higher than the relative position in the \_\_\_\_\_\_\_\_\_ class.

### Percentiles

Percentiles divide a data set into 100 equal groups. The *p*th percentile is a value for which about *p*% of the data values are less than and for which about (100 – *p*)% of the data values are larger than.

Thus, where *p* = percentile

*c* = number of data values below a specific data value

and *n* = total number of data values in the data set, the formula to find a percentile is ; and

the formula to find the data value corresponding to a given percentile is.

If *c* is a not whole number, round up to the next whole number. Then starting at the lowest value, count over to the number that corresponds to the rounded up value. If *c* is a whole number, use the value halfway between the *ct*h and (*c* + 1)st values when counting up from the lowest value.

### Example 3-17. Find Percentiles

A teacher gives a 20-point test to 10 students. The scores are arranged in order from lowest to highest.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

1. Find the percentile for a score of 12.
2. Find the value corresponding to the 60th percentile.

*Solution a:*

c = 7 since 12 is in the 7th position in the ordered list

n = 10

12 is the 75th percentile.

*Solution b:*

n = 10 and p = 60

= 6. 6 is a whole number, so the percentile is the value halfway between the 6th and 7th values when counting up from the lowest value. The 6th value is 10 and the 7th value is 12. Thus, the 60th percentile is \_\_\_\_\_\_.

### Quartiles

Quartiles divide the distribution into four equal groups denoted by *Q*1, *Q*2, and *Q*3, representing the 25th, 50th or median, and 75th percentiles.

To find Data Values Corresponding to *Q*1, *Q*2, and *Q*3.

1. Arrange the data set in order from lowest to highest.
2. Find the median of the data values. This is the value for *Q*2.
3. Find the median of the data values that fall below the median. This is *Q*1.
4. Find the median of the data values that fall above the median. This is *Q*3.

Or use technology.

### Example 3-18. Find the Quartiles for a Data Set

Find the data values corresponding to *Q*1, *Q*2, and *Q*3.

12, 16, 27, 18, 13, 19, 36, 15, 20

*Solution:*

Put the values in order:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Find the median, *Q*2. The median is \_\_\_\_\_\_\_\_\_\_\_\_.

List the values less than the median. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Find the first quartile, *Q*1. *Q*1 = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

List the values greater than the median. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Find the third quartile, *Q*3. *Q*3 = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

### Example 3-19. Find the Quartiles

Find the data values corresponding to *Q*1, *Q*2, and *Q*3

14, 16, 17, 18, 19, 20, 24, 31, 32, 54

*Solution:*

Put the values in order:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Find the median, *Q*2. The median is \_\_\_\_\_\_\_\_\_\_\_\_.

List the values less than the median. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Find the first quartile, *Q*1. *Q*1 = \_\_\_\_\_\_\_\_\_\_\_\_\_

List the values that are greater than the median: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Find the third quartile, *Q*3. *Q*3 = \_\_\_\_\_\_\_\_\_\_\_\_

### Interquartile Range

The Interquartile Range (IQR) is the difference between the third and the first quartiles.

### Example 3-20. Find the Interquartile Range

Find the IQR for the data sets listed in previous two examples.

Data Set 1: 12, 13, 15, 16, 18, 19, 20, 27, 36

Data Set 2: 14, 16, 17, 18, 19, 20, 24, 31, 32, 54

*Solution:*

The IQR for the first data set is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

The IQR for the second data set is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

### Outliers

An outlier is an extremely \_\_\_\_\_\_\_\_ or extremely \_\_\_\_\_\_\_ data value when compared to the rest of the data values.

The mean and standard deviation can be affected strongly by \_\_\_\_\_\_\_\_\_\_\_\_.

Thus, they are nonresistant statistics. The median and interquartile range are less affected by outliers and are called \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ statistics.

To Identify Outliers:

1. Arrange data in order from \_\_\_\_\_\_\_\_\_\_\_\_ to \_\_\_\_\_\_\_\_\_\_\_ .
2. Find the Interquartile range:
3. Multiply the by .
4. Find the upper outlier boundary:

Find the lower outlier boundary:

1. Check the data set for any data value not between the boundaries.

### Example 3-21. Find Outliers

Check the data set for outliers.

14, 16, 17, 18, 19, 20, 24, 31, 32, 54

*Solution:*

\_\_\_\_\_.

The interquartile range \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The lower outlier boundary \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The upper outlier boundary \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Outliers are the data values smaller than \_\_\_\_\_\_ and greater than \_\_\_\_\_. Thus, \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# 3 – 4 Exploratory Data Analysis

Objective 4. Use the Techniques of Exploratory Data Analysis, Including Boxplots and Five-Number Summaries, to Discover Various Aspects of Data.

### Five-Number Summary and Boxplots

The Five-Number Summary includes five specific values:

1. The minimum value of the data set.
2. \_\_\_\_\_\_\_\_\_
3. The maximum value of the data set.

A boxplot is a graphical representation of the data set.

### To Construct a Boxplot

1. Find the five-number summary.
2. Draw a horizontal axis and place the scale on the axis.  
    The scale should start on or below the minimum data value and end on   
    or above the maximum data value.
3. Locate and mark each of the values of the five-number summary. Then  
    draw a box with vertical sides through and . Draw a vertical line   
    through the median. Finally, draw a horizontal line from the minimum   
    data value to the left side of the box and draw a line from the maximum   
    data value to the right side of the box.

### Information Obtained from a Boxplot

1. a. If the median is near the center of the box, the distribution is

approximately symmetric.

b. If the median is to the left of the center of the box, the distribution is

positively skewed.

c. If the median is to the right of the center of the box, the distribution

is negatively skewed.

2. a. If the lines from the box to the minimum and maximum are about

the same length, the distribution is approximately symmetric.

b. If the line from the box to the minimum is longer than the line from

the box to the maximum, the distribution is positively skewed.

c. If the line from the box to the minimum is shorter than the line from

the box to the maximum, the distribution is negatively skewed.

### Example 3-22. Find the Five-Number Summary and Construct a Boxplot

For the given data set used in Example 3-17, find a) the five-number summary and b) construct a boxplot.

12, 16, 27, 18, 13, 19, 36, 15, 20

*Solution:*

* 1. Find the five number summary.

Arrange the data in order: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Find the quartiles (see the previous examples).

The five number summary is:

Maximum = \_\_\_\_\_\_\_

= \_\_\_\_\_\_\_

= Median = \_\_\_\_\_\_\_

= \_\_\_\_\_\_\_

Maximum = \_\_\_\_\_\_\_

* 1. Construct the boxplot.

The boxplot for Example 3-22 has a \_\_\_\_\_\_\_\_\_\_\_\_ line on the left than on the right and the median is to the \_\_\_\_\_\_\_\_\_\_of the center of the box, so the distribution appears to be \_\_\_\_\_\_\_\_\_\_\_\_\_\_ skewed.