Chapter 3 Numerically Summarizing Data

Section 3.1 Measures of Central Tendency

Objectives

1. Determine the arithmetic mean of a variable from raw data
2. Determine the median of a variable from raw data
3. Explain what it means for a statistic to be resistant
4. Determine the mode of a variable from raw data

1 Determine the Arithmetic Mean of a Variable from Raw Data

The **arithmetic mean** of a variable is computed by adding all the values of the variable in the data set and dividing by the number of observations. The **population arithmetic mean**, μ (pronounced "mew"), is computed using all the individuals in a population. The population mean is a parameter.

The **sample arithmetic mean**, $\bar{x}$ (pronounced "x-bar"), is computed using sample data. The sample mean is a statistic.

If $x_1, x_2, \ldots, x_N$ are the N observations of a variable from a population, then the population mean, μ, is

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum x_i}{N} \quad \textbf{(1)}$$

If $x_1, x_2, \ldots, x_n$ are n observations of a variable from a sample, then the sample mean, $\bar{x}$, is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x_i}{n} \quad \textbf{(2)}$$

Note: We use N to represent the size of the population and n to represent the size of the sample. The symbol $\sum$ (the Greek letter capital sigma) tells us to add the terms.

EXAMPLE 1 Computing a Population Mean and a Sample Mean

The following data represent the travel times (in minutes) to work for all seven employees of a start-up web development company.

23, 36, 23, 18, 5, 26, 43

(a) Compute the population mean of this data.
(b) Then take a simple random sample of $n = 3$ employees. Compute the sample mean. Obtain a second simple random sample of $n = 3$ employees. Again compute the sample mean.

Alternative Example 1

Open the data set "Chicago_Salaries" in StatCrunch.

(a) Use StatCrunch (or some other software) to compute the population mean of this data.

(b) Then take a simple random sample of $n = 15$ employees. Compute the sample mean. Obtain a second simple random sample of $n = 15$ employees. Again compute the sample mean.

2 Determine the Median of a Variable from Raw Data

The **median** of a variable is the value that lies in the middle of the data when arranged in ascending order. We use M to represent the median.

Steps in Finding the Median of a Data Set

Step 1 Arrange the data in ascending order.

Step 2 Determine the number of observations, n.

Step 3 Determine the observation in the middle of the data set.

- If the number of observations is odd, then the median is the data value exactly in the middle of the data set. That is, the median is the observation that lies in the $\frac{n+1}{2}$ position.
- If the number of observations is even, then the median is the mean of the two middle observations in the data set. That is, the median is the mean of the observations that lie in the $\frac{n}{2}$ position and the $\frac{n}{2}+1$ position.

Example 2 Determining the Median of a Data Set with an Odd Number of Observations

The following data represent the travel times (in minutes) to work for all seven employees of a start-up web development company.

23, 36, 23, 18, 5, 26, 43

Determine the median of this data.

Example 3 Determining the Median of a Data Set with an Even Number of Observations

Suppose the start-up company hires a new employee. The travel time of the new employee is 70 minutes. Determine the median of the "new" data set.

23, 36, 23, 18, 5, 26, 43, 70

3 Explain What It Means for a Statistic to Be Resistant

Load the Mean Versus Median Applet that is located at www.pearsonhighered.com/sullivanstats . Or, from StatCrunch, select Applets > Mean/SD vs. Median/IQR . Select the "Randomly generated" radio button. Verify the Mean and Median boxes are checked and click Compute!.

1. Click "Reset" at the top of the applet.

a. Create a data set of ten observations such that the mean and median are both roughly equal to 2.

b. Click "Add point" and add a new observation at 9. How does this new value affect the mean? The median?

2. a. Remove the single value near 9 by clicking on the point and dragging it off the number line.

b. Click "Add point" and add a single observation at 24. How does this new value affect the mean? The median?

3. Click "Reset" at the top of the applet.

a. Add a point at 0. Add a second point at 50. Remove these points by dragging them off the screen. (This is done to create a number line from 0 to 50).

b. Create a symmetric data set of six observations such that the mean and median are roughly 40.

c. Add a single observation at 35. How does this new value affect the mean? The median?

d. Grab this new point at 35 and drag it toward 0. What happens to the value of the mean?

What happens to the value of the median?

4. Click “Reset” at the top of the applet.
a. Add a point at 0. Add a second point at 50. Remove these points by dragging them off the screen.
b. Add about 25 to 30 points to create a symmetric dot plot such that the values of the mean and median are roughly 40.

c. Add a single observation at 35. How does this new value affect the mean? The median?

d. Grab this new point at 35 and drag it toward 0. What happens to the value of the mean? What happens to the value of the median?

5. Write a paragraph that summarizes what you have learned in this activity about the mean and median. Be sure to include a discussion of the concept of resistance and the role sample size plays in resistance.

6. Click “Reset” at the top of the applet.
a. Add a point at 0. Add a second point at 50. Remove these points by dragging them off the screen.
b. Create a data set of at least ten observations such that the mean equals the median.

What is the shape of the distribution?

c. Create a data set of at least ten observations such that the mean is greater than the median. What is the shape of the distribution?

d. Create a data set of at least ten observations such that the mean is less than the median. What is the shape of the distribution?

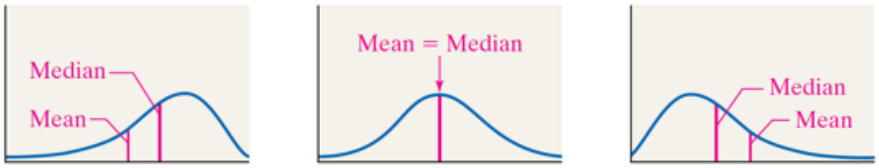
e. Create a data set that is skewed left, with at least 50 observations. Describe the relationship between the mean and the median.

f. Create a data set that is skewed right, with at least 50 observations. Describe the relationship between the mean and the median.

Table 4

Relation between the Mean, Median, and Distribution Shape

Distribution Shape	Mean versus Median
Skewed left	Mean substantially smaller than median
Symmetric	Mean roughly equal to median
Skewed right	Mean substantially larger than median



(a) Skewed Left
Mean < Median

(b) Symmetric
Mean = Median

(c) Skewed Right
Mean > Median

A word of caution is in order. The relation between the mean, median, and skewness are guidelines. The guidelines tend to hold up well for continuous data, but when the data are discrete, the rules can be easily violated.

4 Determine the Mode of a Variable from Raw Data

The **mode** of a variable is the most frequent observation of the variable that occurs in the data set.

To compute the mode, tally the number of observations that occur for each data value. The data value that occurs most often is the mode. A set of data can have no mode, one mode, or more than one mode. If no observation occurs more than once, we say the data have no mode.

Summary: Measures of Central Tendency

Measure of Central Tendency	Computation	Interpretation	When to Use
Mean	Population mean: $\mu = \frac{\Sigma x_i}{N}$ Sample mean: $\bar{x} = \frac{\Sigma x_i}{n}$	Center of gravity	When data are quantitative and the frequency distribution is roughly symmetric
Median	Arrange data in ascending order and divide the data set in half	Divides the bottom 50% of the data from the top 50%	When the data are quantitative and the frequency distribution is skewed left or right
Mode	Tally data to determine most frequent observation	Most frequent observation	When the most frequent observation is the desired measure of central tendency or the data are qualitative

Section 3.2 Measures of Dispersion

Objectives

1. Determine the range of a variable from raw data
2. Determine the standard deviation of a variable from raw data
3. Determine the variance of a variable from raw data
4. Use the Empirical Rule to describe data that are bell shaped
5. Use Chebyshev's Inequality to describe any data set

To order food at a McDonald's restaurant, one must choose from multiple lines, while at Wendy's Restaurant, one enters a single line. The following data represent the wait time (in minutes) in line for a simple random sample of 30 customers at each restaurant during the lunch hour. This data is available in StatCrunch (filename: Wendys vs McDonalds).

Wait Time at Wendy's					
1.50	0.79	1.01	1.66	0.94	0.67
2.53	1.20	1.46	0.89	0.95	0.90
1.88	2.94	1.40	1.33	1.20	0.84
3.99	1.90	1.00	1.54	0.99	0.35
0.90	1.23	0.92	1.09	1.72	2.00

Wait Time at McDonald's					
3.50	0.00	0.38	0.43	1.82	3.04
0.00	0.26	0.14	0.60	2.33	2.54
1.97	0.71	2.22	4.54	0.80	0.50
0.00	0.28	0.44	1.38	0.92	1.17
3.08	2.75	0.36	3.10	2.19	0.23

For each sample, answer the following:

(a) What was the mean wait time?

(b) Draw a histogram of each restaurant's wait time.

(c) Which restaurant's wait time appears more dispersed? Which line would you prefer to wait in? Why?

1 Determine the Range of a Variable from Raw Data

The **range, *R,*** of a variable is the difference between the largest data value and the smallest data values. That is,

$$\text{Range} = R = \text{Largest Data Value} - \text{Smallest Data Value}$$

Example 1 Finding the Range of a Set of Data

The following data represent the travel times (in minutes) to work for all seven employees of a start-up web development company.

23, 36, 23, 18, 5, 26, 43

Find the range.

2 Determine the Standard Deviation of a Variable from Raw Data

The **population standard deviation** of a variable is the square root of the sum of squared deviations about the population mean divided by the number of observations in the population, N. That is, it is the square root of the mean of the squared deviations about the population mean.

The population standard deviation is symbolically represented by σ (lowercase Greek sigma).

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}} = \sqrt{\frac{\Sigma (x_i - \mu)^2}{N}} \quad \textbf{(1)}$$

where $x_1, x_2, \ldots, x_N$ are the N observations in the population and μ is the population mean.

Example 2 Computing a Population Standard Deviation

The following data represent the travel times (in minutes) to work for all seven employees of a start-up web development company.

23, 36, 23, 18, 5, 26, 43

Compute the population standard deviation of this data.

x_i	μ	$x_i - \mu$	$(x_i - \mu)^2$
23	24.85714	-1.85714	3.44898
36	24.85714	11.14286	124.1633
23	24.85714	-1.85714	3.44898
18	24.85714	-6.85714	47.02041
5	24.85714	-19.8571	394.3061
26	24.85714	1.142857	1.306122
43	24.85714	18.14286	329.1633

$$\sum (x_i - \mu)^2 = 902.8571$$

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}} = \sqrt{\frac{902.8571}{7}} \approx 11.36 \text{ minutes}$$

The **sample standard deviation**, s, of a variable is the square root of the sum of squared deviations about the sample mean divided by $n - 1$, where n is the sample size.

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}} = \sqrt{\frac{\Sigma (x_i - \bar{x})^2}{n - 1}} \quad \textbf{(3)}$$

where $x_1, x_2, \ldots, x_n$ are the n observations in the sample and $\bar{x}$ is the sample mean.

We call $n - 1$ the degrees of freedom because the first $n - 1$ observations have freedom to be whatever value they wish, but the nth value has no freedom. It must be whatever value forces the sum of the deviations about the mean to equal zero.

Example 3 Computing a Sample Standard Deviation

Here are the results of a random sample taken from the travel times (in minutes) to work for all seven employees of a start-up web development company:

5, 26, 36

Find the sample standard deviation.

x_i	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
5	22.33333	-17.333	300.432889
26	22.33333	3.667	13.446889
36	22.33333	13.667	186.786889

$$\sum (x_i - \bar{x})^2 = 500.66667$$

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{500.66667}{2}} \approx 15.82 \text{ minutes}$$

Example 4 Comparing Standard Deviations

Determine the standard deviation waiting time for Wendy's and McDonald's. Which is larger? Why?

The standard deviation may be used to judge whether a particular observation is "far away" from the mean of a data set. For example, is 31 cm far from a mean of 25 cm? If the standard deviation of the data set is 6 cm, then the answer is no because 31 would be only 1 standard deviation from 25. However, if the standard deviation is 2 cm, then the answer is yes because 31 would be 3 standard deviations from 25.

So, when judging the unusualness of an observation, it is vital that you consider the underlying variation in the data as measured by the standard deviation.

Far Away? Is 1000 "far away" from 900. As someone with knowledge of measures of dispersion, you answer should be "it depends." Consider the following.

(a) Let's say the standard deviation is 25. Is 1000 "far away" from 900?

(b) Now let's say the standard deviation is 120. Is 1000 "far away" from 900?

3 Determine the Variance of a Variable from Raw Data

The **variance** of a variable is the square of the standard deviation. The **population variance** is σ^2 and the **sample variance** is s^2.

Example 5 Computing a Population Variance

The following data represent the travel times (in minutes) to work for all seven employees of a start-up web development company.

23, 36, 23, 18, 5, 26, 43

Compute the population and sample variance of this data.

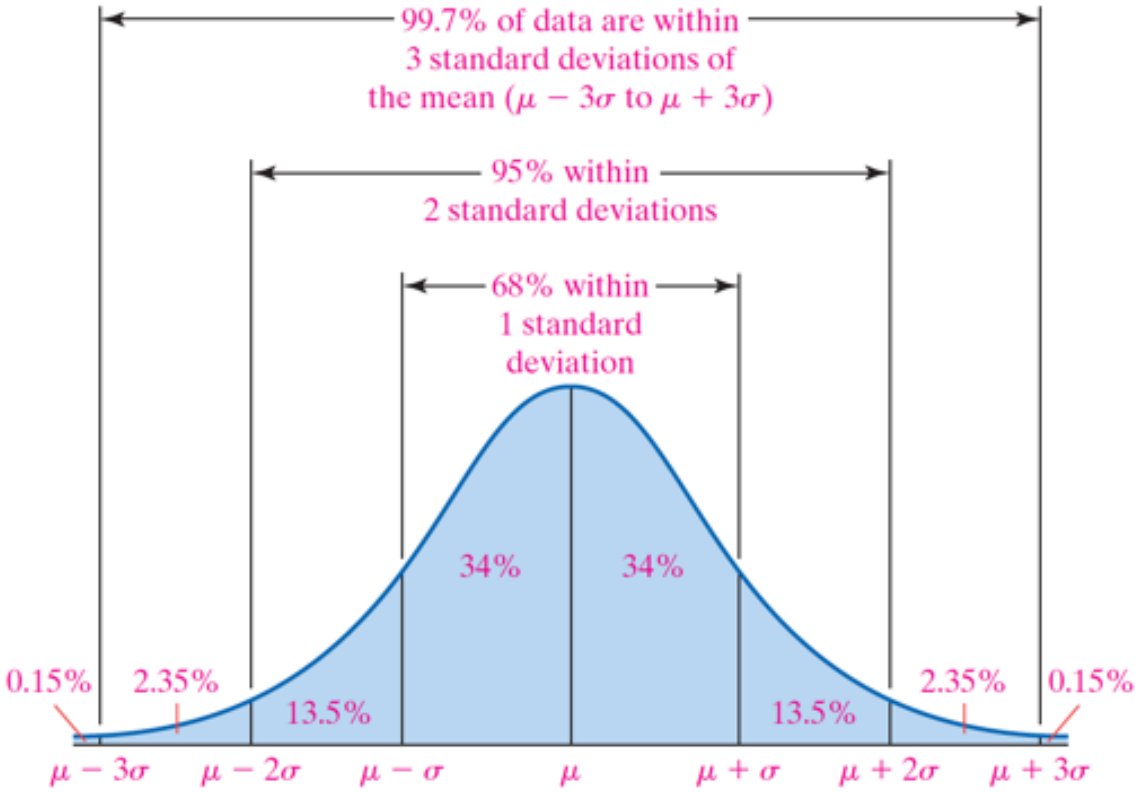
4 Use the Empirical Rule to Describe Data That Are Bell-Shaped

The Empirical Rule

If a distribution is roughly bell shaped, then

- Approximately 68% of the data will lie within 1 standard deviation of the mean. That is, approximately 68% of the data lie between $\mu - 1\sigma$ and $\mu + 1\sigma$.
- Approximately 95% of the data will lie within 2 standard deviations of the mean. That is, approximately 95% of the data lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.
- Approximately 99.7% of the data will lie within 3 standard deviations of the mean. That is, approximately 99.7% of the data lie between $\mu - 3\sigma$ and $\mu + 3\sigma$.

Note: We can also use the Empirical Rule based on sample data with $\bar{x}$ used in place of μ and s used in place of σ.



Example 6 Using the Empirical Rule

The waist circumference of 2-year-old males is bell-shaped with mean 48.5 cm and standard deviation 4.8 cm.
(a) About 95% of 2-year-old males will have waist circumferences between what values?
(b) What percentage of 2-year-old males have waist circumference between 34.1 cm and 62.9 cm?
(c) What percentage of 2-year-old males have waist circumference between 53.3 cm and 62.9 cm?

5 Use Chebyshev's Inequality to Describe Any Set of Data

Chebyshev's Inequality

For any data set or distribution, at least $\left(1 - \frac{1}{k^2}\right) \cdot 100\%$ of the observations lie within k standard deviations of the mean, where k is any number greater than 1. That is, at least $\left(1 - \frac{1}{k^2}\right) \cdot 100\%$ of the data lie between $\mu - k\sigma$ and $\mu + k\sigma$ for $k > 1$.

Note: We can also use Chebyshev's Inequality based on sample data.

Section 3.3 Measures of Central Tendency and Dispersion from Grouped Data

Objectives

1. Approximate the mean of a variable from grouped data
2. Compute the weighted mean
3. Approximate the standard deviation of a variable from grouped data

Approximate the Mean and Standard Deviation of a Variable from Grouped Data

We have discussed how to compute descriptive statistics from raw data, but often the only available data have already been summarized in frequency distributions (**grouped data**). Although we cannot find exact values of the mean or standard deviation without raw data, we can approximate these measures using the techniques discussed in this section.

Approximate Mean of a Variable from a Frequency Distribution

Population Mean

$$\mu = \frac{\sum x_i f_i}{\sum f_i} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots + f_n}$$

Sample Mean

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots + f_n} \quad \textbf{(1)}$$

where x_i is the midpoint or value of the ith class

f_i is the frequency of the ith class

n is the number of classes

Approximate Standard Deviation of a Variable from a Frequency Distribution

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2 f_i}{\sum f_i}}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2 f_i}{\left(\sum f_i\right) - 1}} \quad \textbf{(3)}$$

where x_i is the midpoint or value of the ith class

f_i is the frequency of the ith class

Example 1 Approximate the Mean and Standard Deviation from Grouped Data

A simple random sample of 89 two-year old Toyota Prius cars that are listed for sale was collected from www.cars.com. The advertised prices of the cars are summarized in the table below. Find the approximate mean and standard deviation for the advertised prices of the cars.

Price (in dollars)	15,000-17,499	17,500-19,999	20,000-22,499	22,500-24,999	25,000-27,499	27,500-29,999	30,000-32,499	32,500-34,999
Frequency	1	2	4	27	38	15	0	2

Compute the Weighted Mean

The **weighted mean**, $\bar{x}_w$, of a variable is found by multiplying each value of the variable by its corresponding weight, adding these products, and dividing this sum by the sum of the weights. It can be expressed using the formula

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n} \quad \textbf{(2)}$$

where w_i is the weight of the ith observation

x_i is the value of the ith observation

Example 2 Computing the Weighted Mean

Bob goes to the "Buy the Weigh" Nut store and creates his own bridge mix. He combines 1 pound of raisins, 2 pounds of chocolate covered peanuts, and 1.5 pounds of cashews. The raisins cost $1.25 per pound, the chocolate covered peanuts cost $3.25 per pound, and the cashews cost $5.40 per pound. What is the cost per pound of this mix?

Section 3.4 Measures of Position and Outliers

Objectives

1. Determine and interpret *z*-scores
2. Interpret percentiles
3. Determine and interpret quartiles
4. Determine and interpret the interquartile range
5. Check a set of data for outliers

1 Determine and Interpret *z*-Scores

The ***z*-score** represents the distance that a data value is from the mean in terms of the number of standard deviations. We find it by subtracting the mean from the data value and dividing this result by the standard deviation. There is both a population z-score and a sample z-score:

Population *z*-Score	**Sample *z*-Score**	
$z = \frac{x - \mu}{\sigma}$	$z = \frac{x - \bar{x}}{s}$	**(1)**

The z-score is unitless. It has mean 0 and standard deviation 1.

Example 1 Comparing *z*-Scores

The mean upper arm length of 19-year-old males is 38.6 cm with a standard deviation of 2.9 cm. The mean upper arm length of 19-year-old females is 35.8 cm with a standard deviation of 2.8 cm. Who has a relatively longer upper arm length – a male whose upper arm length is 41.1 cm or a female whose upper arm length is 38.4 cm?

2 Interpret Percentiles

The ***k*th percentile**, denoted P_k, of a set of data is a value such that k percent of the observations are less than or equal to the value.

Example 2 Interpret a Percentile

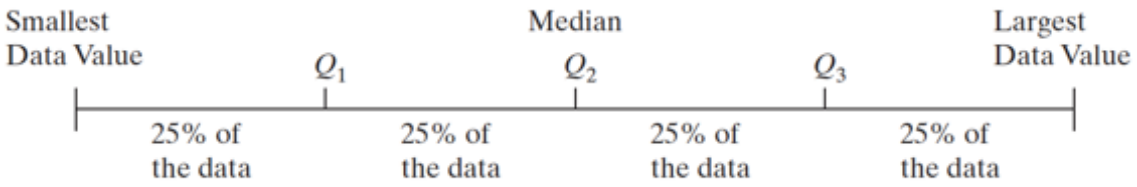
The Graduate Record Examination (GRE) is a test required for admission to many U.S. graduate schools. The University of Pittsburgh Graduate School of Public Health requires a GRE score no less than the 70th percentile for admission into their Human Genetics MPH or MS program. Source: http://www.publichealth.pitt.edu/interior.php?pageID=101

Interpret this admissions requirement.

3 Determine and Interpret Quartiles

Quartiles divide data sets into fourths, or four equal parts.

- The first quartile, denoted Q_1, divides the bottom 25% of the data from the top 75%. Therefore, the first quartile is equivalent to the 25th percentile.
- The second quartile, Q_2, divides the bottom 50% of the data from the top 50%; it is equivalent to the 50th percentile or the median.
- The third quartile, Q_3, divides the bottom 75% of the data from the top 25%; it is equivalent to the 75th percentile.



Finding Quartiles

Step 1 Arrange the data in ascending order.

Step 2 Determine the median, M, or second quartile, Q_2.

Step 3 Divide the data set into halves: the observations below (to the left of) M and the observations above M. The first quartile, Q_1, is the median of the bottom half of the data and the third quartile, Q_3, is the median of the top half of the data.

Example 3 Finding Quartiles

Download the "PayScale_ROI_2017" data from StatCrunch. Determine and interpret the quartiles for ROI (return on investment).

4 Determine and Interpret the Interquartile Range

The **interquartile range, IQR**, is the range of the middle 50% of the observations in a data set. That is, the IQR is the difference between the third and first quartiles and is found using the formula

$$\text{IQR} = Q_3 - Q_1$$

Example 4 Determine and Interpret the Interquartile Range

Find and interpret the interquartile range of the data from Example 3.

Summary: Which Measures to Report

Shape of Distribution	Measure of Central Tendency	Measure of Dispersion
Symmetric	Mean	Standard deviation
Skewed left or skewed right	Median	Interquartile range

5 Check a Set of Data for Outliers

Extreme observations are referred to as **outliers.**

Checking for Outliers by Using Quartiles

Step 1 Determine the first and third quartiles of the data.

Step 2 Compute the interquartile range.

Step 3 Determine the fences. **Fences** serve as cutoff points for determining outliers.

$$\text{Lower fence} = Q_1 - 1.5(\text{IQR})$$
$$\text{Upper fence} = Q_3 + 1.5(\text{IQR})$$

Step 4 If a data value is less than the lower fence or greater than the upper fence, it is considered an outlier.

Example 5 Checking for Outliers

Check the data from Example 3 for outliers.

Section 3.5 The Five-Number Summary and Boxplots

Objectives

1. Compute the five-number summary
2. Draw and interpret boxplots

1 Compute the Five-Number Summary

The five-number summary of a set of data consists of the smallest data value, Q_1, the median, Q_3, and the largest data value. We organize the five-number summary as follows:

Five-Number Summary

MINIMUM	Q_1	M	Q_3	MAXIMUM

Example 1 Computing the Five-Number Summary

Download the "PayScale_ROI_2017" data from StatCrunch. Determine the five-number summary for ROI (return on investment).

2 Draw and Interpret Boxplots

Drawing a Boxplot

Step 1 Determine the lower and upper fences:

$$\text{Lower fence} = Q_1 - 1.5(\text{IQR})$$
$$\text{Upper fence} = Q_3 + 1.5(\text{IQR})$$
$$\text{where IQR} = Q_3 - Q_1$$

Step 2 Draw a number line long enough to include the maximum and minimum values. Insert vertical lines at Q_1, M, and Q_3. Enclose these vertical lines in a box.

Step 3 Label the lower and upper fences.

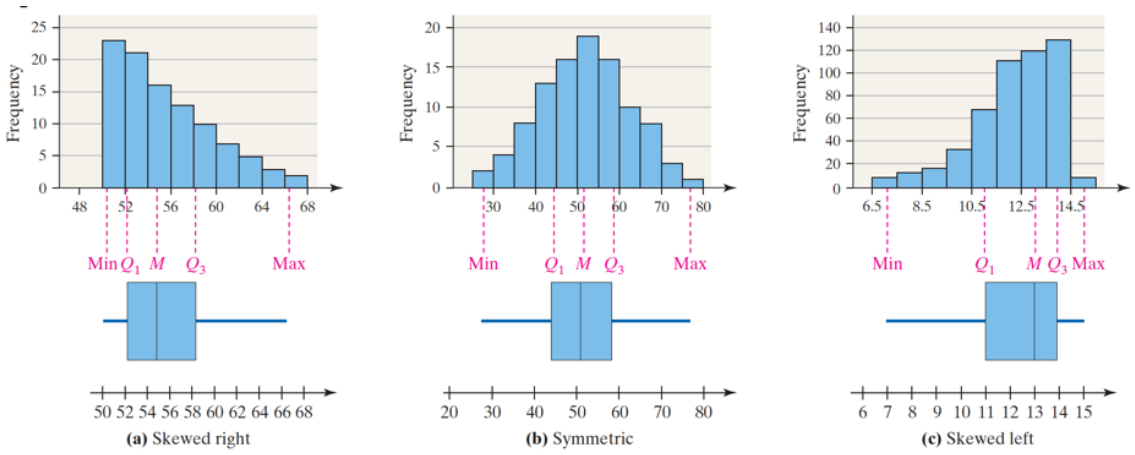
Step 4 Draw a line from Q_1 to the smallest data value that is larger than the lower fence. Draw a line from Q_3 to the largest data value that is smaller than the upper fence. These lines are called **whiskers**.

Step 5 Any data values less than the lower fence or greater than the upper fence are outliers and are marked with an asterisk (*).

Example 2 Constructing a Boxplot

Download the "PayScale_ROI_2017" data from StatCrunch. Construct a boxplot for ROI (return on investment).

Using a Boxplot and Quartiles to Describe the Shape of a Distribution



Example 4 Comparing Two Distributions Using Boxplots

Draw side-by-side boxplots of the Wendys versus McDonalds data from Section 3.2.