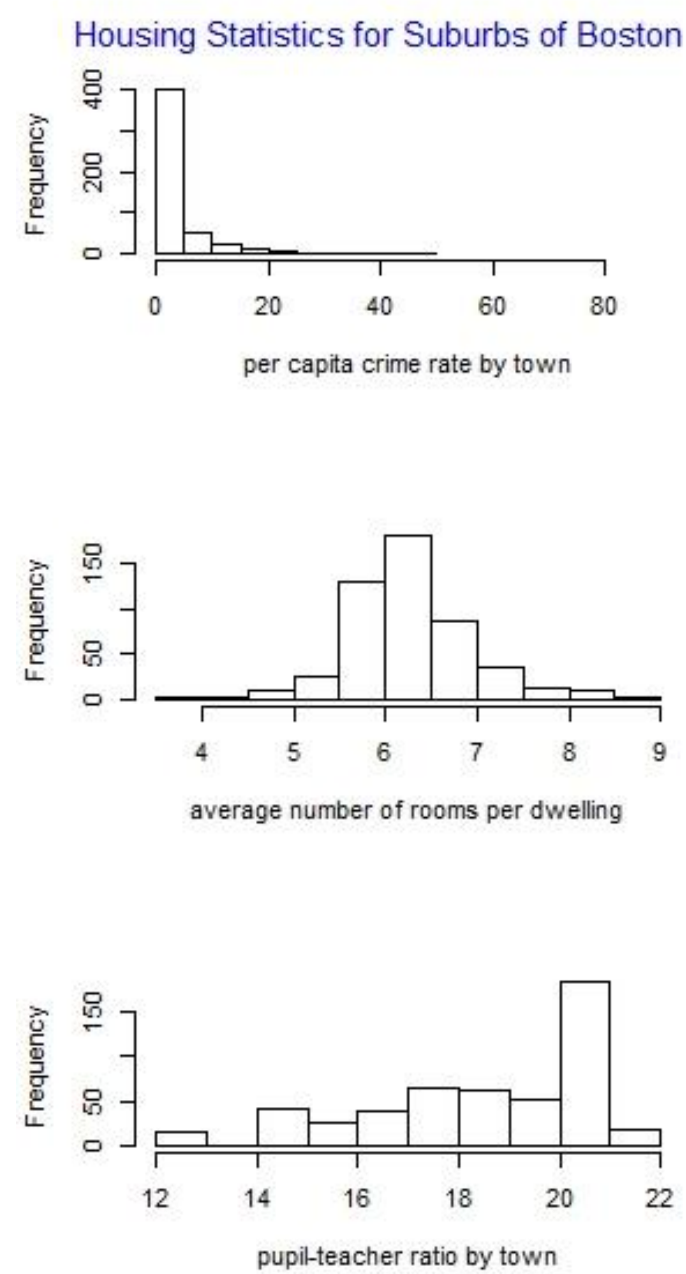
CHAPTER 3: Data Description

You've tabulated and made pretty pictures. Now what numbers do you use to summarize your data?

You’ll find a link on our website to a data set with various measures for housing in the suburbs of Boston. It comes from a paper titled: “Hedonic prices and the demand for clean air.” I’ve given histograms here for a few of the variables.

- **What are some characteristics of the distributions that you might want to describe?**

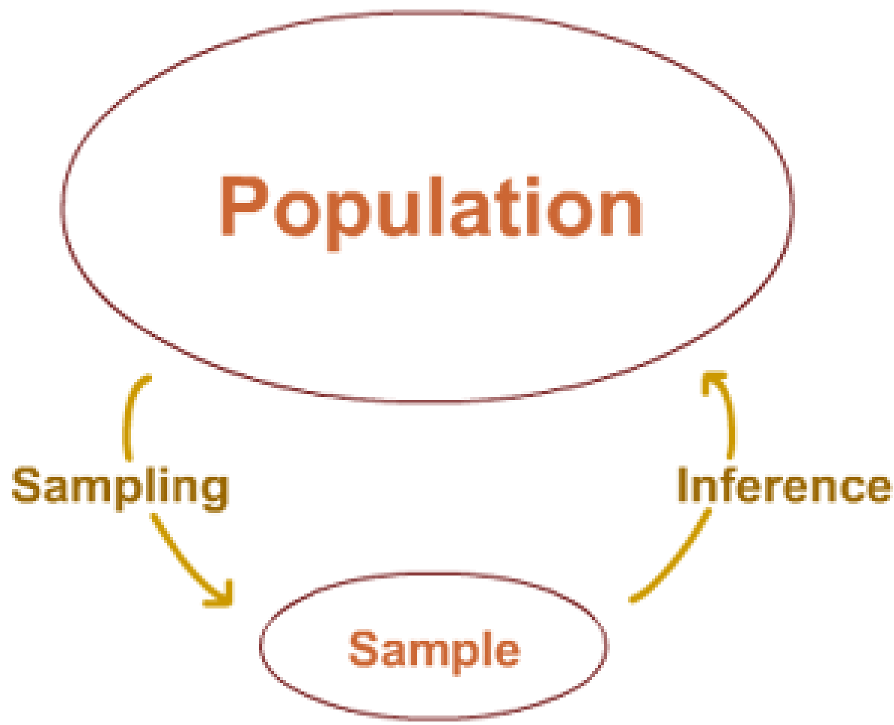
- **Do you think these measures might do a better job for some of these variables versus others? Why or why not?**



What's being described?

Parameter – a characteristic or measure obtained using the data values from a specific **population**.

Statistic – a characteristic or measure obtained using the data values from a **sample**.



Rules and Notation:

- Let x represent the variable for which we have sample data.
- Let n represent the number of observations in the sample. (the sample size).
- Let N represent the number of observations in the population.
- $\sum x$ represents the sum of all the data values of x.
- $\sum x^2$ is the sum of the data values after squaring them.
- $\left(\sum x\right)^2 \neq \sum x^2$.

General Rounding Rule: When computations are done in the calculation, rounding should not be done until the final answer is calculated!

Rounding Rule of Thumb for Calculations from Raw Data: The final answer should be rounded to one more decimal place than that of the original data. You will see that this will be true for the mean, variance and standard deviation.

Section 3-1: Measures of Central Tendency

Measure	**Description**	**Statistic and Parameter**	**Notes and Insights**
Mean	the sum of the data values divided by the total number of values	The sample mean is denoted by $\overline{X}$ and calculated using the formula: $\overline{x} = \frac{\sum x}{n}$ The population mean is denoted by μ and is found with the formula: $\mu = \frac{\sum x}{N}$	The mean should be rounded to one more decimal place than occurs in the raw data. The mean is the balance point of the data. When the data is skewed the mean is pulled in the direction of the longer tail. The mean is used in computing other statistics such as variance and standard deviation. The mean is highly affected by outliers and may not be an appropriate statistic to use when an outlier is present.
Median	the middle number of the data set when they are ordered from smallest to largest	Arrange the data in order. If *n* is odd, the median is the middle number. If *n* is even, the median is the mean of the middle two numbers We use the symbol MD for median.	The median is robust against outliers (less affected by them). The median is used when one must find the center value of a data set
Mode	the value that occurs most often in a data set		This is where the "peaks" occur in a histogram. **Unimodal** – when a data set has only one mode **Bimodal –** when a data set has 2 modes **Multimodal –** when a data set has more than 2 modes **No Mode** – when no data values occurs more than once The mode is used when the most typical case is desired.

Measures of Central Tendency continued…

Measure	**Description**	**Statistic and Parameter**	**Notes and Insights**
Midrange	the sum of the minimum and maximum values in the data set, divided by 2	Denoted by the symbol MR. $\text{MR} = \frac{\min + \max}{2}$	It is a *rough estimate* of the middle since it gives the midpoint of the dataset. An outlier (a really large or really small value) can have a dramatic effect on the midrange value.
Weighted Mean	found by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights.	$\bar{x} = \frac{w_1x_1 + w_2x_2 + \ldots + w_nx_n}{w_1 + w_2 + \ldots + w_n}$ where $w_1, w_2, \ldots, w_n$ are the weights and $x_1, x_2, \ldots, x_n$ are the 1^{st}, 2^{nd}, … , *n*th data values.	

GROUP WORK: (use appropriate notation)

Find the Mean, Median and Midrange of the daily vehicle pass charge for five U.S. National Parks. The costs are $25, $15, $15, $20, and $25.

Find the Mean, Median and Midrange of the numbers of water-line breaks per month in the last two winter seasons for the city of Brownsville, Minnesota: 2, 3, 6, 8, 4, 1. Find the midrange.

Find the mode of the following data sets:

Set 1: 12, 8, 14, 15, 11, 10, 5, 14
Set 2: 1, 2, 3, 4
Set 3: 1, 2, 3, 4, 1, 2, 3
Set 4: 18.0, 14.0, 34.5, 10, 11.3, 10, 12.4, 10

Find the weighted mean for the grade point using the number of credits for a class as the weight:

Course	Credits (weights)	Grade point(x)
English	3	C (2 points)
Calculus	4	A (4 points)
Yoga	2	A (4 points)
Physics	4	B (3 points)

Now consider which of these measures would be good representations of "central tendency" for the 3 variables from the Boston housing data set.

	Per Capita Crime Rate By Town	**Average Number Of Rooms Per Dwelling**	**Pupil-Teacher Ratio By Town**
$\bar{x} =$	3.613524	6.2846	18.46
MD=	0.256510	6.2085	19.05
MR=	44.491260	6.1705	17.30
Mode=	0.01501, 14.3337 (each occurring twice)	5.713, 6.127, 6.167, 6.229, 6.405, 6.417 (all occurring 3 times)	20.2 (occurring 140 times; the next count closest occurred 34 times)

Notice how the statistics compare to each other for each variable, e.g., mean, median and midrange are all close to each other for the room variable. Why? Why is this not the case for the other variables?

Location of Mean, Median, and Mode on Distribution Shapes

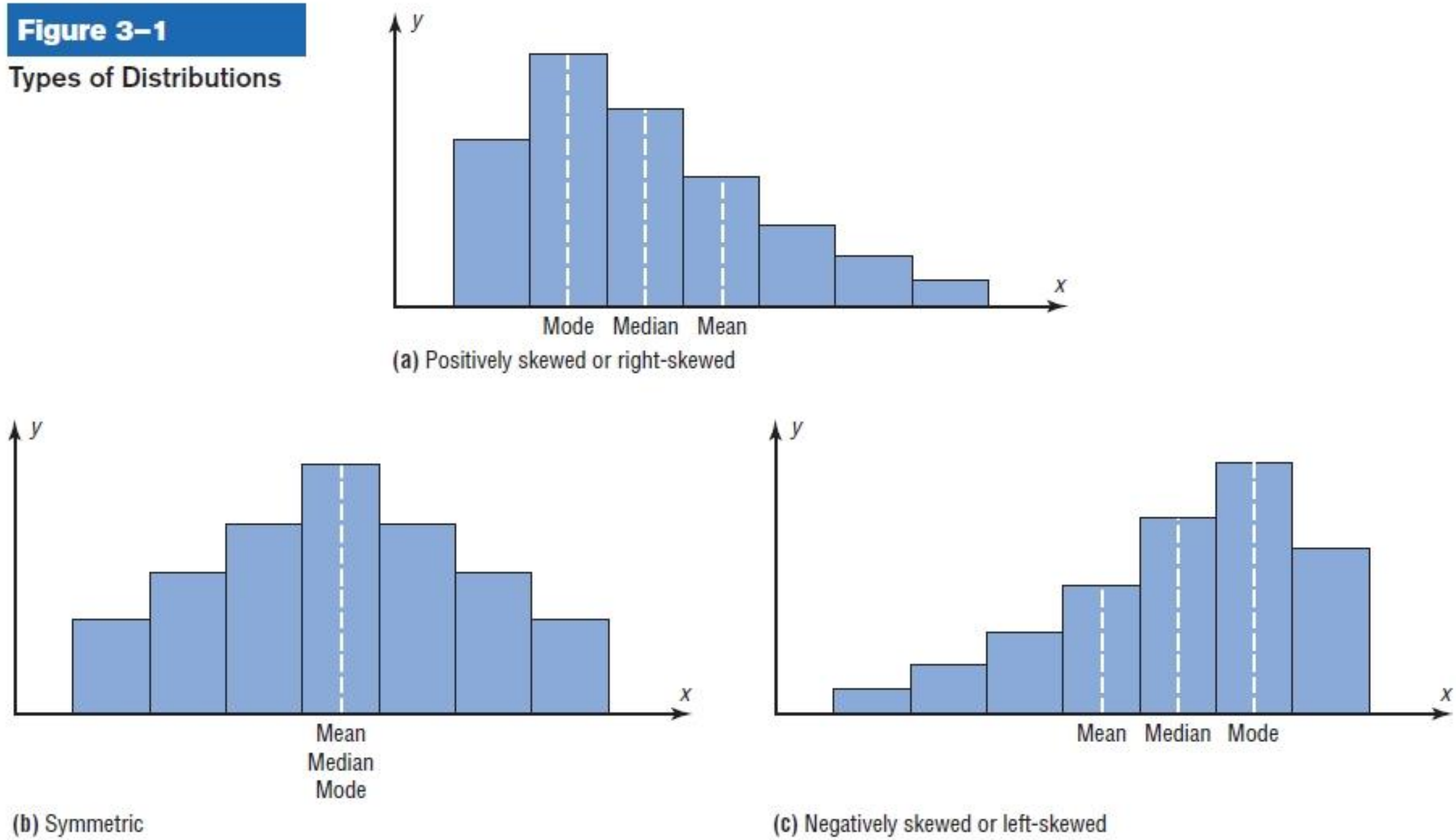


Figure 3–1 Types of Distributions

Section 3-2: Measures of Variation (Spread)

To describe a distribution well, we need more than just the measures of center. We also need to know how the data is spread out or how it varies.

Example: Bowling scores

John	185	135	200	185	250	155
Jarrod	182	185	188	185	180	190

Who is the better bowler and how do you know?

Note: The mean score for both bowlers is 185.

Ways to Measure Spread:

Measure	Description	Sample	Population	Notes
Range	The difference between the largest and smallest observations.	Denoted by *R*. *R* = high value – low value		
Variance	The average of the squares of the distance each value is from the mean	The sample variance is an estimate of the population variance calculated from a sample. It is denoted by s^2. The formula to calculate the sample variance is $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$. or $s^2 = \frac{n\left(\sum x^2\right)-\left(\sum x\right)^2}{n(n-1)}$	Population variance is denoted by σ^2 It is commonly used in statistics because it has nice theoretical properties. The formula for the population variance is $\sigma^2 = \frac{\sum(x-\mu)^2}{N}$.	In practice, we don't know the population values or parameters, so we cannot calculate σ^2 or σ. We end up calculating the variance and standard deviation of a sample. Be careful to notice the difference of n-1 (sample) and n (population) in the denominator.
Standard deviation	the "typical" deviation from the sample mean	The square root of the sample variance It is denoted by *s*. The formula to calculate the sample standard deviation is: $s = \sqrt{s^2} = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$. OR $s = \sqrt{\frac{n\left(\sum x^2\right)-\left(\sum x\right)^2}{n(n-1)}}$	the square root of the population variance The symbol for the population standard deviation is σ The formula for the population standard deviation is $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x-\mu)^2}{N}}$.	The greater the spread of the data, the larger the value of *s*. *s*=0 only when all observations take the same value. *s* can be influenced by outliers because outliers influence the mean and because outliers have large deviations from the mean

Steps for Calculating Sample Variance and Standard Deviation

1. Calculate the sample mean $\overline{X}$.
2. Calculate the deviation from the mean for every data value (data value – mean).
3. Square all the values from #2 and find the sum.
4. Divide the sum in #3 by $n-1$. This calculation produces the **sample variance.**
5. Take the square root of #4. This number produces the **sample standard deviation**.

The same (general) procedure applies for finding the population variance and standard deviation except we use the population mean μ and divide by N instead of $n-1$.

Group work: Compute the range, sample variance and sample standard deviation for Jarrod's bowling scores. If this was our whole population, what will differ? For the sake of comparison, John's statistics are: $\bar{x} = 185,\ s^2 = 1570,\ s = 39.6$.

And to check your answer, John's statistics are:

$\bar{x} = 185,\ s^2 = 13.6,\ s = 3.7$

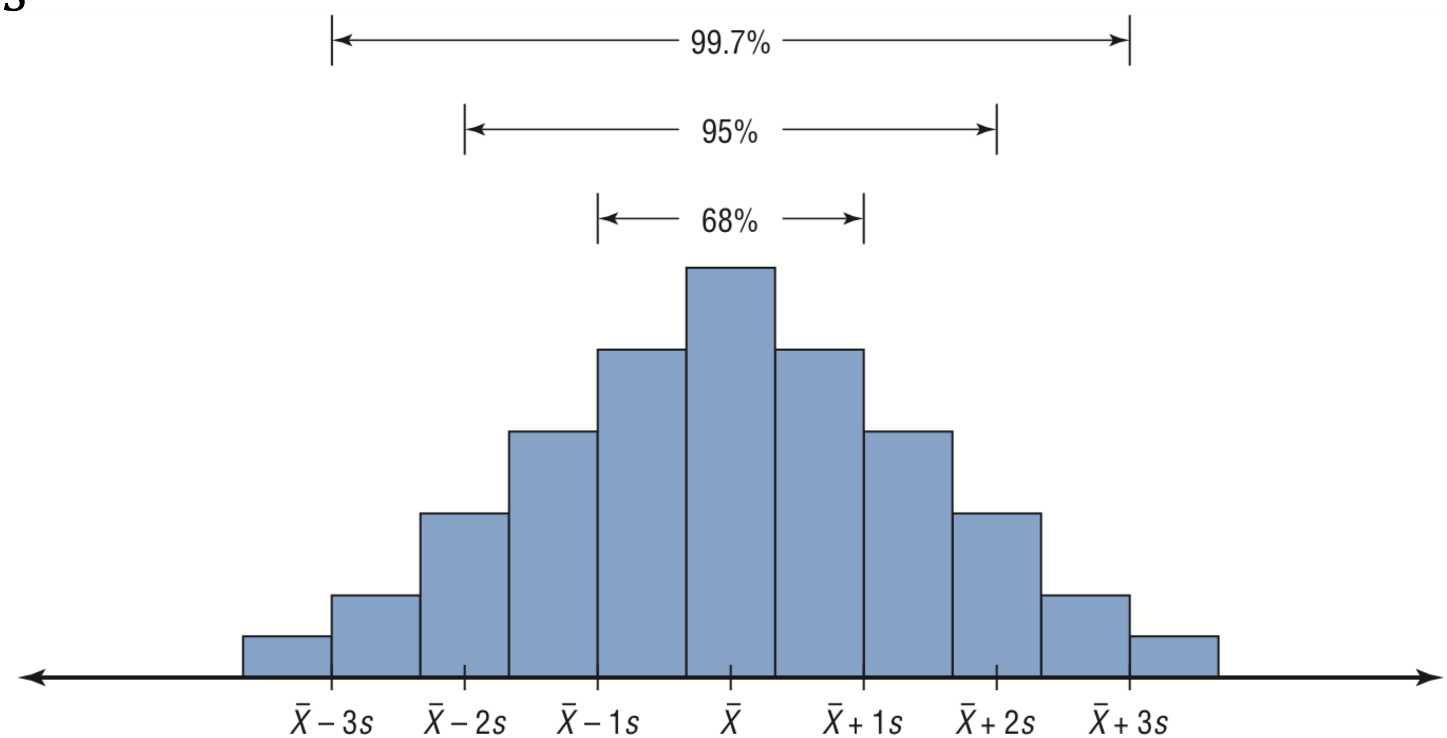
Uses of the Variance and Standard Deviation

1. To determine the spread of data. The larger the variance or standard deviation, the greater the data are dispersed.
2. Makes it easy to compare the dispersion of two or more data sets to decide which is more spread out.
3. To determine the consistency of a variable. E.g., the variation of nuts and bolts in manufacturing must be small.
4. Frequently used in inferential statistics, as we will see later in the book.
5. Empirical rule……

Empirical Rule

Works ONLY for symmetric, unimodal curves (bell-shaped curves):

- Approximately 68% of data falls within 1 standard deviation of the mean, $\bar{x} \pm s$
- Approximately 95% of data falls within 2 standard deviations of the mean, $\bar{x} \pm 2s$
- Approximately 99.7% of data falls within 3 standard deviations of the mean, $\bar{x} \pm 3s$



Example: Mothers' Heights

An article in 1903 published the heights of 1052 mothers. The sample mean was 62.484 inches and the standard deviation was 2.390 inches.

Note the summary table below regarding the actual percentages and the empirical rule.

Number of sd's	Interval	Actual	Empirical Rule	Chebyshev Rule
1	60.094 to 64.874	72.1%	68%	At least 0%
2	57.704 to 67.264	96.2%	95%	At least 75%
3	55.314 to 69.654	99.2%	99.7%	At least 89%

Section 3-3: Measure of Position
(some of…this section we need for use in Section 3-4)

Quartiles – values that divide the distribution into four groups, separated by Q1, Q2 (median), and Q3.

- Q1 is the 25th percentile.
- Q2 is the 50th percentile (the median).
- Q3 is the 75th percentile.

Interquartile Range (IQR) – the difference between Q1 and Q3. This is the range of the middle 50% of the data.

$$\text{IQR} = \text{Q3} - \text{Q1}$$

Finding the Quartiles:

1. Arrange the data from smallest to largest.
2. Find the median data value. This is the value for Q2.
3. Find the median of the data values that fall below Q2. This is the value for Q1. (Don't include the median in these values if the number of observations is odd.)
4. Find the median of the data values that are greater than Q2. This is the value for Q3. (Don't include the median in these values if the number of observations is odd.)

Example: Find Q1, Q2, and Q3 for the data set 15, 13, 6, 5, 12, 50, 22, 18. Additionally, what is the IQR?

Q2 (*Median) is (13+15)/2 = 14*

Order data: 5, 6, 12, 13, 15, 18, 22, 50

Q1 is (6+12)/2 =9

Q3 is (18+22)/2=20

IQR = Q3 – Q1 = 20 – 9 = 11

Outlier – an extremely large or small data value when compared to the rest of the data values.

- One way to identify outliers is by defining an outlier to be any data value that has value more than 1.5 times the IQR from Q1 or Q3.
- i.e., a data value is an outlier if it is smaller than $Q1-1.5(IQR)$ or larger than $Q3+1.5(IQR)$.

Steps to Identify Outliers:

1. Arrange the data in order and find Q1, Q3, and the IQR.
2. Calculate Q1 – 1.5(IQR) and Q3 + 1.5(IQR).
3. Find any data values smaller than the Q1 – 1.5(IQR) or larger than Q3 + 1.5(IQR).

Example: Find the outliers of the data set 15, 13, 6, 5, 12, 50, 22, 18. Note: This is the same data set as the previous example.

Step 1: Q1 = 9, Q3 = 20, IQR = 11

Step 2: Q1 – 1.5(IQR) = 9 – 1.5(11) = -7.5
Q3 + 1.5(IQR) = 20 + 1.5(11) = 36.5

Step 3: The data value, 50, is considered an outlier.

Section 3-4: Exploratory Data Analysis

Exploratory Data Analysis (EDA) - Examining data to find out what information can be discovered about the data such as the center and the spread.

- Uses robust statistics such as the median and interquartile range.

Five number summary – the minimum, Q1, median (Q2), Q3, and the maximum of a data set.

Boxplot – a graphical display of the five-number summary (and potentially outliers) using a "box" and "whiskers".

- Outliers are indicated by * on the boxplot.

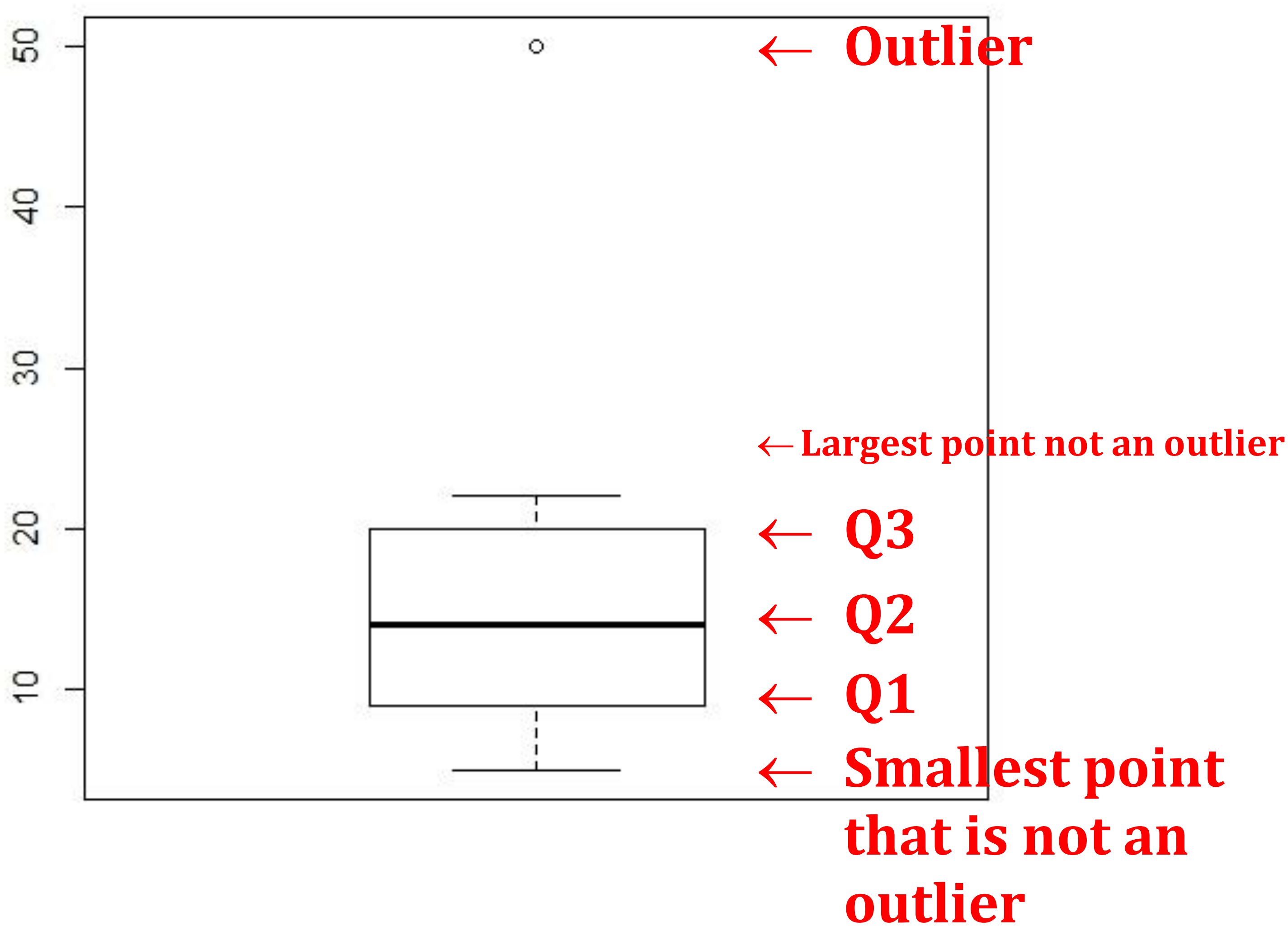
Steps for constructing a boxplot:

1. Order the data and calculate the five-number summary.
2. Determine if there are any outliers.
3. Draw and scale an axis (either horizontal or vertical) that includes both the minimum and the maximum.
4. Draw a box extending from Q1 to Q3.
5. Draw a bar across the box at the value of the median.
6. Draw bars extending from away from the box extending to the most extreme values that are not outliers.
7. Draw stars for all of the outliers.

Note: There are many slight variations of boxplots. The book's version of a boxplot doesn't mark outliers, just the max and min with bars. **The book calls our boxplot a "modified boxplot".**

Example: Create a boxplot for the data set 15, 13, 6, 5, 12, 50, 22, 18. Note: This is the same data set as the previous example.

Min = 5
Q1 = 9
Q2 = 14
Q3 = 20
Max = 50
Outlier at 50



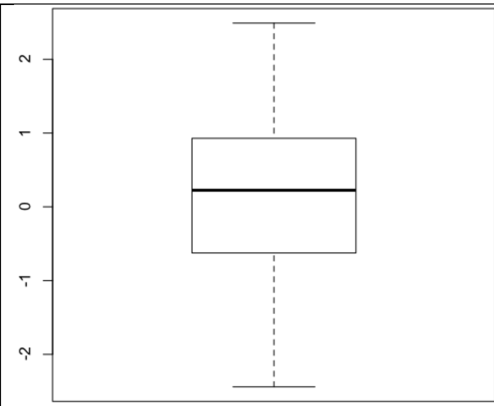
Example: Draw a boxplot for the following data set:
2, 5, 5, 7, 7, 8, 9, 10, 10, 10, 10, 14, 17, 20

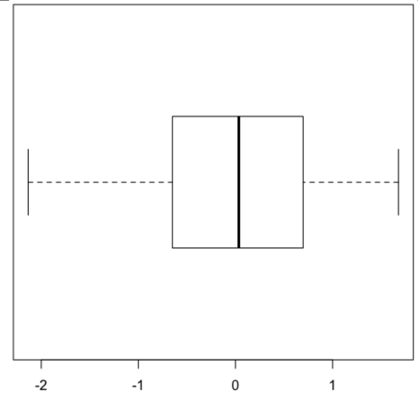
Q_1=7, Median=9.5, Q_3=10

In general, a box plot gives us the following information:

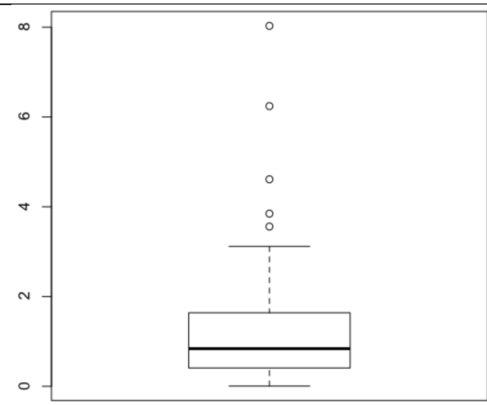
- If the median is near the center of the box and the bars are about the same length, the distribution is approximately symmetric.
- If the median is lower than the "middle" of the box and the bar going in the positive direction is larger than the one in the negative direction, then the distribution is positively-skewed (right-skewed).
- If the median is higher than the "middle" of the box and the bar going in the negative direction is larger than the one in the positive direction, then the distribution is negatively-skewed (left-skewed).

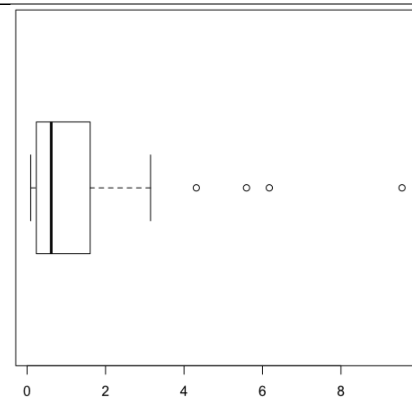
Symmetric Distribution





Positively-skewed (Right-skewed)





Negatively-skewed (Left-skewed)