Chapter 1: Exploring Data

- **Introduction**: **Data Analysis: Making Sense of Data**
- **1.1** Analyzing Categorical Data
- **1.2** Displaying Quantitative Data with Graphs
- **1.3** Describing Quantitative Data with Numbers

Source: The Practice of Statistics, 4th edition - For AP* by STARNES, YATES, MOORE

Introduction
Data Analysis: Making Sense of Data

- **Statistics** is the science of data.
  - **Data Analysis** is the process of *organizing*, *displaying*, *summarizing*, and *asking questions* about data.

Definitions:

Individuals – objects (people, animals, things) described by a set of data

Distribution – tells us what values a variable takes and how often it takes those values

Variable - any characteristic of an individual

Categorical Variable
– places an individual into one of several groups or categories.
– **GRAPHS: Bar and Pie Charts**

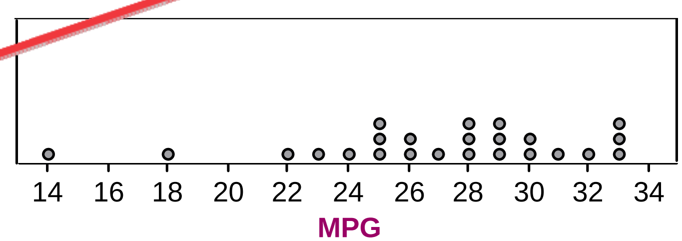
Quantitative Variable
– takes numerical values for which it makes sense to find an average.
– **GRAPHS: Dot Plots, Histograms, Stem-Leaf, and Box Plots**

How to Explore Data

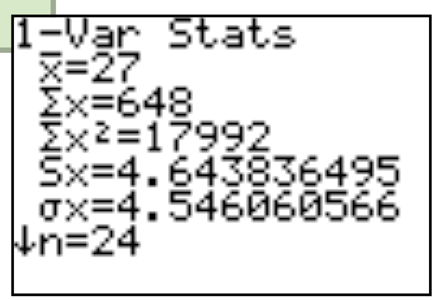
Examine each variable by itself. Then study relationships among the variables.

MODEL	MPG	MODEL	MPG	MODEL	MPG
Acura RL	22	Dodge Avenger	30	Mercedes-Benz E350	24
Audi A6 Quattro	23	Hyundai Elantra	33	Mercury Milan	29
Bentley Arnage	14	Jaguar XF	25	Mitsubishi Galant	27
BMW 5281	28	Kia Optima	32	Nissan Maxima	26
Buick Lacrosse	28	Lexus GS 350	26	Rolls Royce Phantom	18
Cadillac CTS	25	Lincoln MKZ	28	Saturn Aura	33
Chevrolet Malibu	33	Mazda 6	29	Toyota Camry	31
Chrysler Sebring	30	Mercedes-Benz E350	24	Volkswagen Passat	29

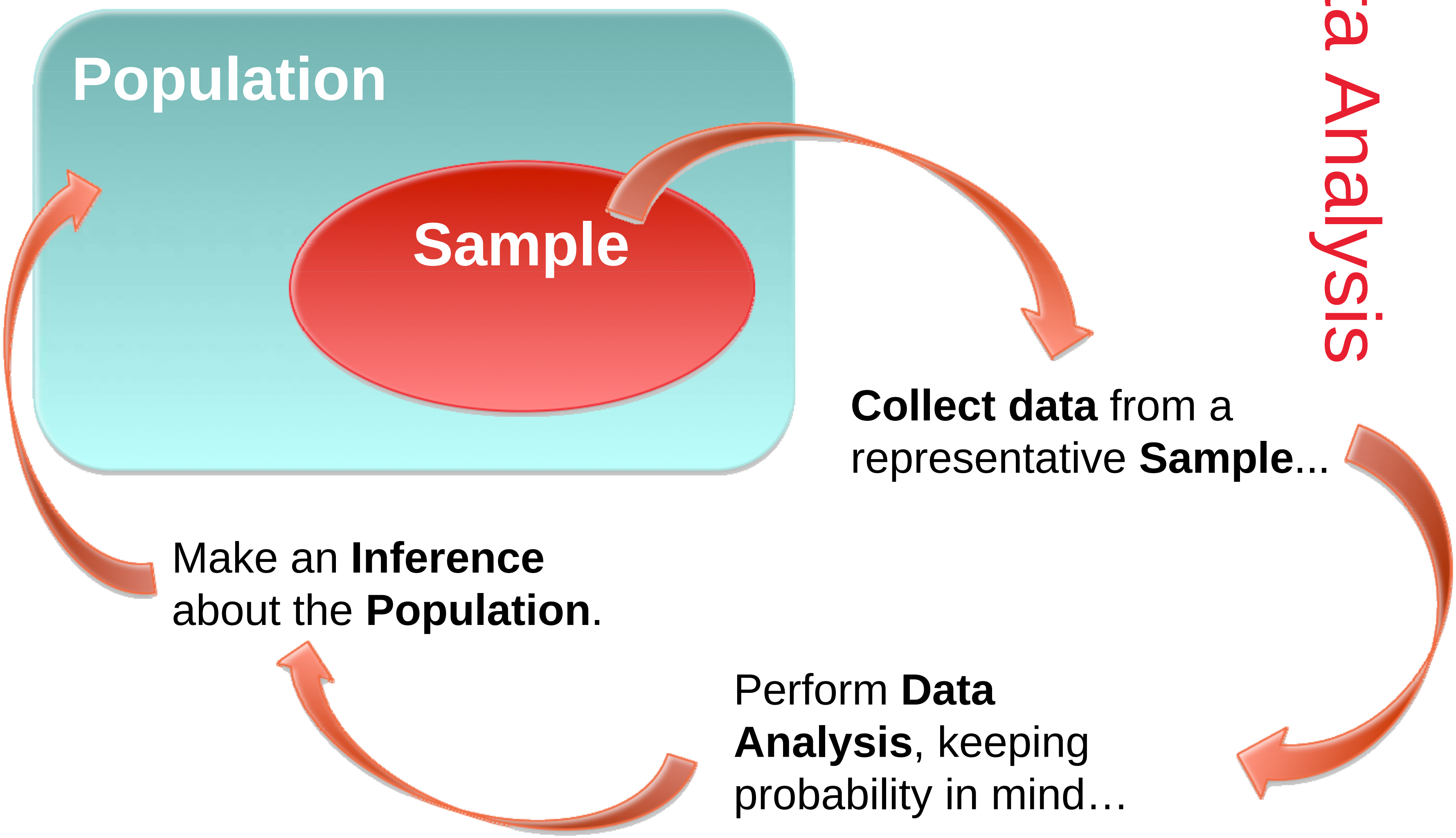
Start with a graph or graphs



Add numerical summaries


1-Var Stats
$\bar{x}=27$
$\Sigma x=648$
$\Sigma x^2=17992$
$Sx=4.643836495$
$\sigma x=4.546060566$
↓n=24

From Data Analysis to Inference



Section 1.1 Analyzing Categorical Data

- **Categorical Variables** place individuals into one of several groups or categories
  - **The values of a categorical variable are labels for the different categories**
  - The distribution of a categorical variable lists the count or percent of individuals who fall into each category.

Example, page 8

Variable

Values

Count

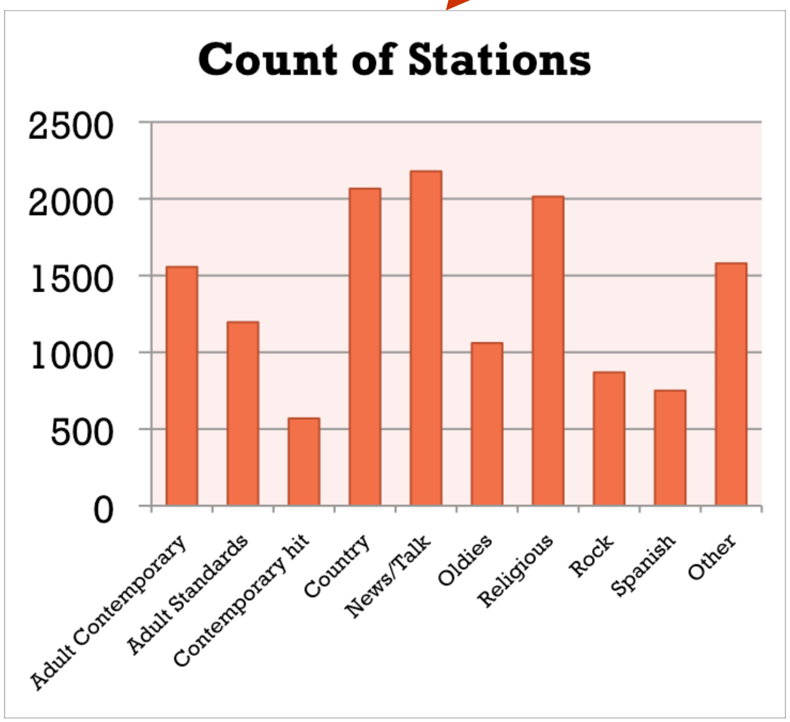
Percent

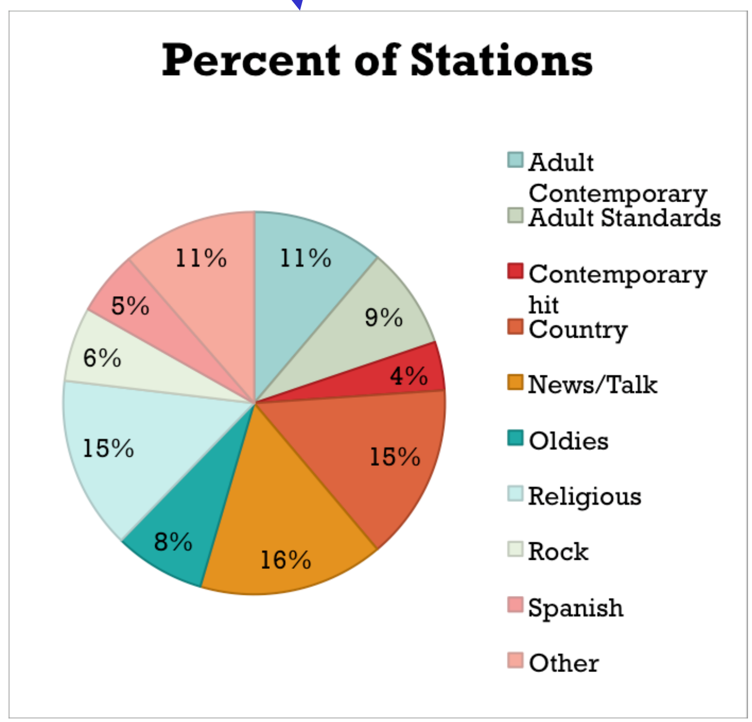
Frequency Table	
Format	**Count of Stations**
Adult Contemporary	1556
Adult Standards	1196
Contemporary Hit	569
Country	2066
News/Talk	2179
Oldies	1060
Religious	2014
Rock	869
Spanish Language	750
Other Formats	1579
Total	**13838**

Relative Frequency Table	
Format	**Percent of Stations**
Adult Contemporary	11.2
Adult Standards	8.6
Contemporary Hit	4.1
Country	14.9
News/Talk	15.7
Oldies	7.
Religious	14.6
Rock	6.3
nguage	5.4
Other Form	11.4
Total	**99.9**

+ Displaying categorical data

- Frequency tables can be difficult to read.
- Sometimes is is easier to analyze a distribution by displaying it with a **bar graph (COUNTS)** or **pie chart (PERCENTS)**.





+ Graphs: Good and Bad

- **Bar graphs compare several quantities by comparing the heights of bars that represent those quantities.**
- **Our eyes react to the *area* of the bars as well as height. Be sure to make your bars equally wide.**
- **Avoid the temptation to replace the bars with pictures for greater appeal…this can be misleading!**

Alternate Example

This ad for DIRECTV has multiple problems.

•First, the heights of the bars are not accurate.

•According to the graph, the difference between 81 and 95 is much greater than the difference between 56 and 81.

•Also, the extra width for the DIRECTV bar is deceptive since our eyes respond to the area, not just the height.

Marginal Distributions in Two-Way Tables

- When a dataset involves two categorical variables, we begin by examining the counts or percents in various categories for *one* of the variables.

Definition:

Two-way Table – describes two categorical variables, organizing counts according to a *row variable* and a *column variable*.

Example, p. 12

Young adults by gender and chance of getting rich			
Opinion	Female	Male	**Total**
Almost no chance	96	98	194
Some chance, but probably not	426	286	712
A 50-50 chance	696	720	1416
A good chance	663	758	1421
Almost certain	486	597	1083
Total	2367	2459	4826

- **What are the variables described by this two-way table?**
- **How many young adults were surveyed?**
- **How many females surveyed?**

Two-Way Tables and Marginal Distributions

Definition:

The **Marginal Distribution** of one of the categorical variables in a two-way table of counts is the distribution of values of that variable among all individuals described by the table.

Note: Percents are often more informative than counts, especially when comparing groups of different sizes.

To examine a marginal distribution,

1) Marginal distribution (in %'s) are the row or column percents.
   What % are female? 2367/4836=.4895 ≈ 49%
2) Make a graph to display the marginal distribution.

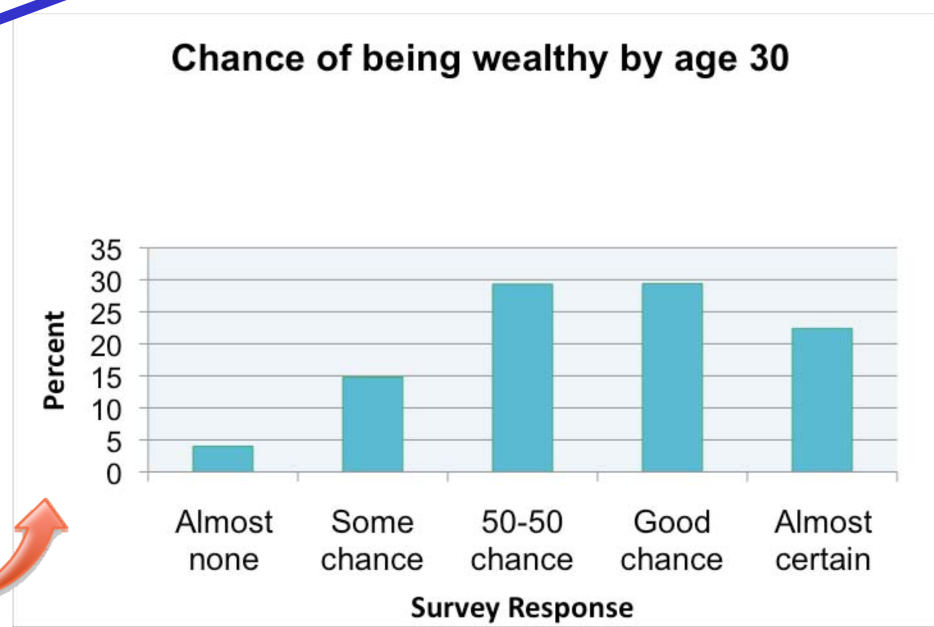
+ Two-Way Tables and Marginal Distributions

Example, p. 13

Young adults by gender and chance of getting rich			
	Female	Male	Total
Almost no chance	96	98	194
Some chance, but probably not	426	286	712
A 50-50 chance	696	720	1416
A good chance	663	758	1421
Almost certain	486	597	1083
Total	2367	2459	4826

Examine: **marginal distribution of chance of getting rich.**

Response	Percent
Almost no chance	194/4826 = **4.0%**
Some chance	712/4826 = **14.8%**
A 50-50 chance	1416/4826 = **29.3%**
A good chance	1421/4826 = **29.4%**
Almost certain	1083/4826 = **22.4%**



Conditional Distributions:

To examine the Relationships Between Categorical Variables

- **Note: Marginal distributions do not tell us anything about the relationship between two variables.**

Definition:

A **Conditional Distribution** of a variable describes the values of that variable among individuals who have a specific value of another variable.

To examine or compare conditional distributions,

1) Select the row(s) or column(s) of interest.
2) Use the data in the table to calculate the conditional distribution (in percents) of the row(s) or column(s).
3) Make a graph to display the conditional distribution.
   - Use a **side-by-side bar graph** or **segmented bar graph** to compare distributions.

Conditional Distributions in Two-Way Tables

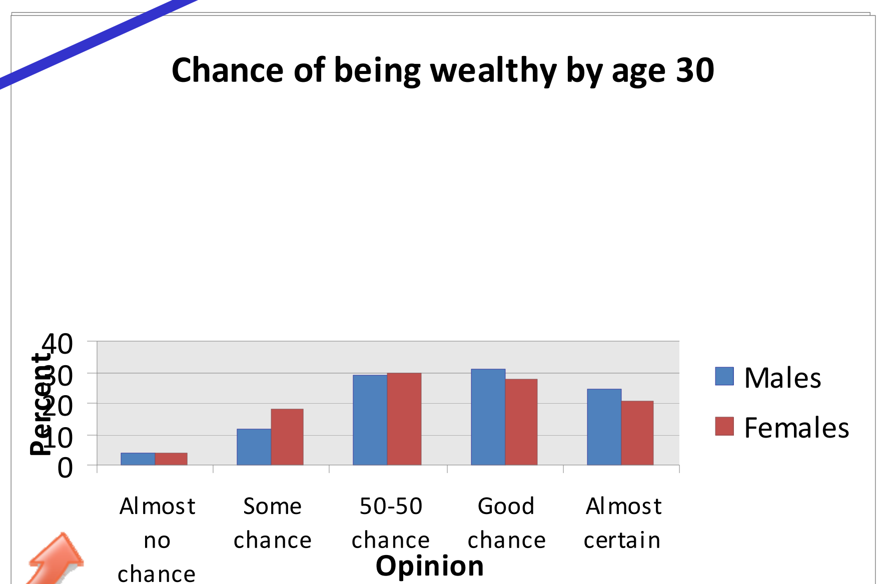
Example, p. 15

Young adults by gender a… …ting rich			
	Female	Male	Total
Almost no chance	96	98	194
Some chance, but probably not	426	286	712
A 50-50 chance	696	720	1416
A good chance	663	758	1421
Almost certain	486	597	1083
Total	2367	2459	4826

Examine the relationship between gender and opinion.

- **Calculate the conditional distribution of opinion among males… then females**

Response	Male	Female
Almost no chance	98/2459 = **4.0%**	96/2367 = **4.1%**
Some chance	286/2459 = **11.6%**	426/2367 = **18.0%**
A 50-50 chance	720/2459 = **29.3%**	696/2367 = **29.4%**
A good chance	758/2459 = **30.8%**	663/2367 = **28.0%**
Almost certain	597/2459 = **24.3%**	486/2367 = **20.5%**



Section 1.2 - Displaying Quantitative Data with Graphs

- **Examining the Distribution of a Quantitative Variable**

- The purpose of a graph is to help us understand the data. After you make a graph, always ask, “What do I see?”

How to Examine the Distribution of a Quantitative Variable

In any graph, look for the **overall pattern** and for striking **departures** from that pattern.

Describe the overall pattern of a distribution by its:

- **Shape**
- **Center**
- **Spread**

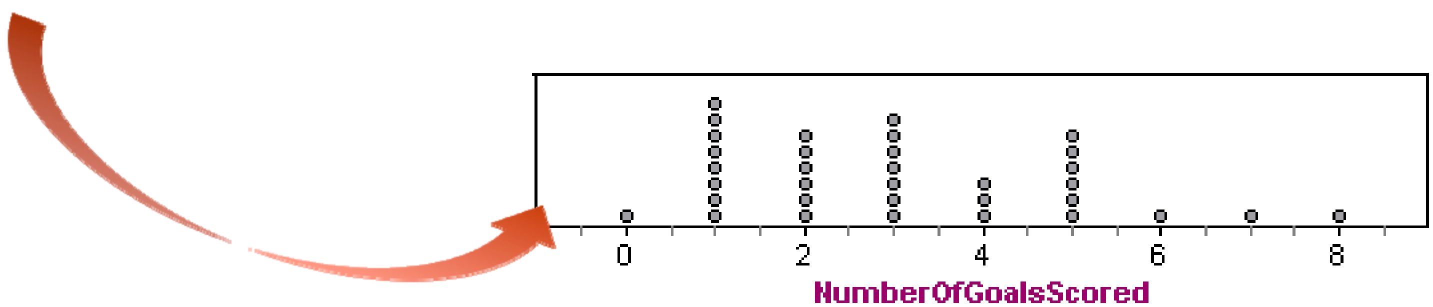
Don’t forget your SOCS! Or CUSS and BS!!

Note individual values that fall outside the overall pattern. These departures are called **outliers**.

Dotplots

- One of the simplest graphs to construct and interpret is a **dotplot**.
- Each data value is shown as a dot above its location on a number line.

Number of Goals Scored Per Game by the 2004 US Women's Soccer Team																
3	0	2	7	8	2	4	3	5	1	1	4	5	3	1	1	3
3	3	2	1	2	2	2	4	3	5	6	1	5	5	1	1	5



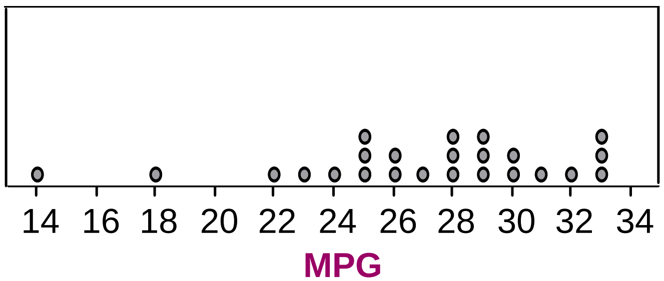
+

CUSS and BS

Example, page 28

- The table and dotplot below displays the Environmental Protection Agency's estimates of highway gas mileage in miles per gallon (MPG) for a sample of 24 model year 2009 midsize cars.

MODEL	MPG	MODEL	MPG	MODEL	MPG
Acura RL	22	Dodge Avenger	30	Mercedes-Benz E350	24
Audi A6 Quattro	23	Hyundai Elantra	33	Mercury Milan	29
Bentley Arnage	14	Jaguar XF	25	Mitsubishi Galant	27
BMW 5281	28	Kia Optima	32	Nissan Maxima	26
Buick Lacrosse	28	Lexus GS 350	26	Rolls Royce Phantom	18
Cadillac CTS	25	Lincoln MKZ	28	Saturn Aura	33
Chevrolet Malibu	33	Mazda 6	29	Toyota Camry	31
Chrysler Sebring	30	Mercedes-Benz E350	24	Volkswagen Passat	29



Describe the shape, center, and spread of the distribution. Are there any outliers?

Describing Shape

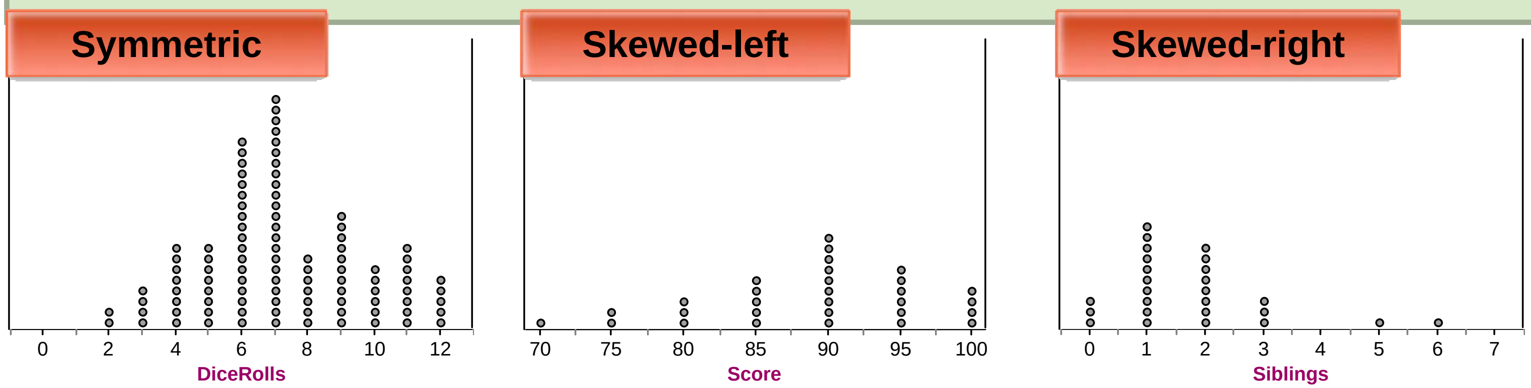
- When you describe a distribution's shape, concentrate on the main features. Look for rough **symmetry** or clear **skewness**.

Definitions:

A distribution is roughly **symmetric** if the right and left sides of the graph are approximately mirror images of each other.

A distribution is **skewed to the right** (right-skewed) if the right side of the graph (containing the half of the observations with larger values) is much longer than the left side.

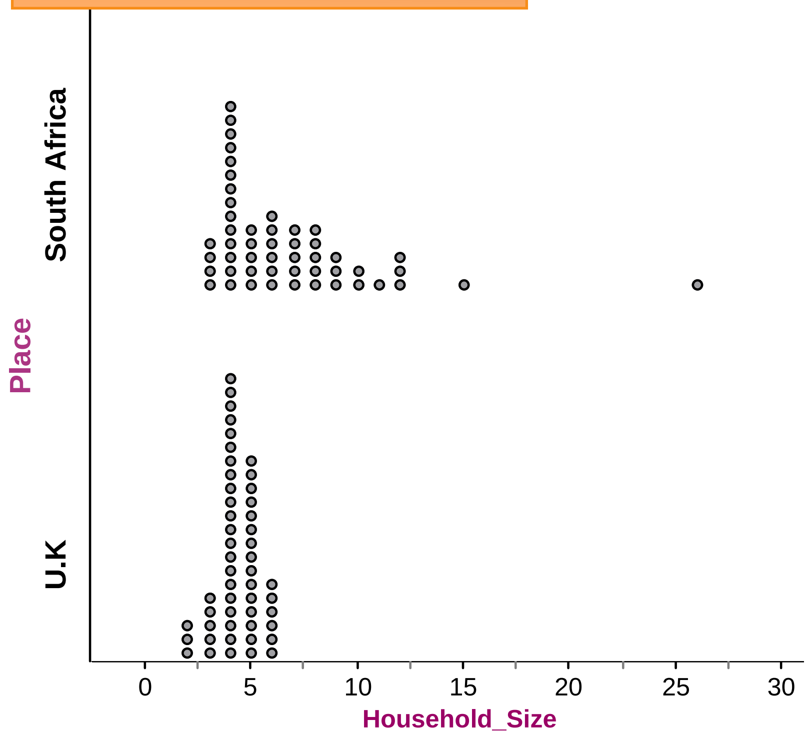
It is **skewed to the left** (left-skewed) if the left side of the graph is much longer than the right side.



Comparing Distributions

- Some of the most interesting statistics questions involve comparing two or more groups.
- Always discuss shape, center, spread, and possible outliers whenever you compare distributions of a quantitative variable.

Example, page 32



Compare the distributions of household size for these two countries. Don't forget your SOCS!

Stemplots (Stem-and-Leaf Plots)

- Another simple graphical display for small data sets is a stemplot.
- Stemplots give us a quick picture of the distribution while including the actual numerical values.

- ***These data represent the responses of 20 female AP Statistics students to the question, "How many pairs of shoes do you have?" Construct a stemplot.***

50	26	26	31	57	19	24	22	23	38
13	50	13	34	23	30	49	13	15	51

Stems		Add leaves		Order leaves	
1		1	93335	1	33359
2		2	664233	2	233466
3		3	1840	3	0148
4		4	9	4	9
5		5	0701	5	0017

Key: 4|9 =49

Or

Key:

Stem=tens

Leaf=ones

Stems **Add leaves** **Order leaves** **Add a key**

Splitting Stems and Back-to-Back Stemplots

- When data values are "bunched up", we can get a better picture of the distribution by **splitting stems**.
- Two distributions of the same quantitative variable can be compared using a **back-to-back stemplot** with common stems.

Females

50	26	26	31	57	19	24	22	23	38
13	50	13	34	23	30	49	13	15	51

Males

14	7	6	5	12	38	8	7	10	10
10	11	4	5	22	7	5	10	35	7

0L
0H
1L
1H
2L
2H
3L
3H
4L
4H
5L
5H

"split stems"

Females		Males
	0	4
	0	555677778
333	1	0000124
95	1	
4332	2	2
66	2	
410	3	
8	3	58
	4	
9	4	
100	5	
7	5	

Key: 4|9 = 49

Histograms

- Quantitative variables often take many values. A graph of the distribution may be clearer if nearby values are grouped together.
- The most common graph of the distribution of one quantitative variable is a **histogram**.

How to Make a Histogram

1)Divide the range of data into classes of equal width.

2)Find the count (*frequency*) or percent (*relative frequency*) of individuals in each class.

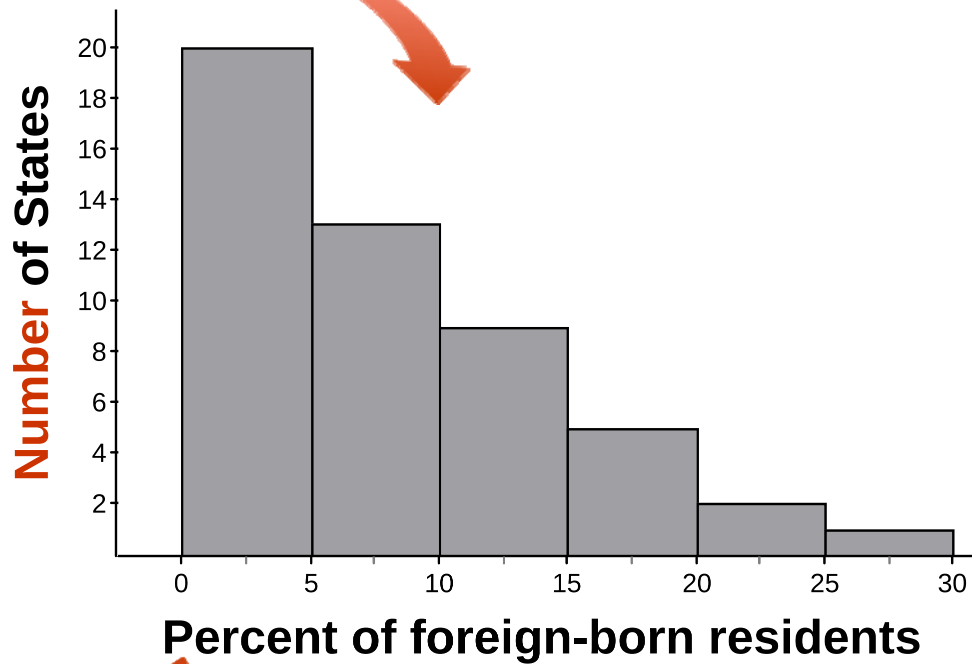
3)Label and scale your axes and draw the histogram. The height of the bar equals its frequency. Adjacent bars should touch, unless a class contains no individuals.

Example, page 35

Making a Histogram

- The table on page 35 presents data on the percent of residents from each state who were born outside of the U.S.

Frequency Table	
Class	**Count**
0 to <5	20
5 to <10	13
10 to <15	9
15 to <20	5
20 to <25	2
25 to <30	1
Total	50



Using Histograms Wisely

- Here are several cautions based on common mistakes students make when using histograms.

Cautions

1)Don't confuse *histograms* and *bar graphs.*

2)Don't use counts (in a frequency table) or percents (in a relative frequency table) as data.

3)Use percents instead of counts on the vertical axis when comparing distributions with different numbers of observations.

4)Just because a graph looks nice, it's not necessarily a meaningful display of data.

Section 1.3
Describing Quantitative Data with Numbers

Learning Objectives

After this section, you should be able to…

- ✓ MEASURE center with the mean and median
- ✓ MEASURE spread with standard deviation and interquartile range
- ✓ IDENTIFY outliers
- ✓ CONSTRUCT a boxplot using the five-number summary
- ✓ CALCULATE numerical summaries with technology

+ Measuring Center: The Mean

- The most common measure of center is the ordinary arithmetic average, or **mean**.

Definition:

To find the **mean** $\bar{x}$ (pronounced "**x-bar**") of a set of observations, add their values and divide by the number of observations (**n**). The observations are $x_1, x_2, x_3, \ldots, x_n$, their mean is:

$$\bar{x} = \frac{\text{sum of observations}}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

In mathematics, the capital Greek letter Σ is short for "summation or simply add them all up." Therefore, the formula for the mean can be written in more compact notation:

$$\bar{x} = \frac{\sum x_i}{n}$$

Measuring Center: The Median

- Another common measure of center is the **median**.
- The median describes the midpoint of a distribution.

Definition:

The **median M** is the midpoint of a distribution, the number such that half of the observations are smaller and the other half are larger.

To find the median of a distribution:

1)Arrange all observations from smallest to largest.

2)If the number of observations *n* is odd, the median *M* is the center observation in the ordered list.

3)If the number of observations *n* is even, the median *M* is the average of the two center observations in the ordered list.

Measuring Center

- Use the data below to calculate the mean and median of the commuting times (in minutes) of 20 randomly selected New York workers.

Example, page 53

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

$$\bar{x} = \frac{10 + 30 + 5 + 25 + \ldots + 40 + 45}{20} = 31.25 \text{ minutes}$$

0	5
1	005555
2	00**05**
3	00
4	005
5	
6	005
7	
8	5

Key: 4|5 =45

$$M = \frac{20 + 25}{2} = 22.5 \text{ minutes}$$

Comparing the Mean and the Median

- The mean and median measure center in different ways, and both are useful.
  - *Don't confuse* ***the "average" value*** *of a variable (****the mean****) with*
  - *its* ***"typical" value****, which we might describe by* ***the median****.*

Comparing the Mean and the Median

The mean and median of a roughly symmetric distribution are close together.

If the distribution is **exactly symmetric, the mean and median are exactly the same.**

In a **skewed distribution,** the mean is usually farther out in the long tail than is the median. That is the mean is pulled towards the outliers – **skewed left or skewed right.**

Measuring Spread: The Interquartile Range (*IQR*)

- A measure of center alone can be misleading.
- A useful numerical description of a distribution requires both a measure of center and a measure of spread.

How to Calculate the Quartiles and the Interquartile Range

To calculate the **quartiles**:

1)Arrange the observations in increasing order and locate the median *M*.

2)The **first quartile** Q_1 is the median of the observations located to the left of the median in the ordered list.

3)The **third quartile** Q_3 is the median of the observations located to the right of the median in the ordered list.

The **interquartile range (*IQR*)** is defined as:

$$IQR = Q_3 - Q_1$$

Find and Interpret the IQR

Example, page 57

Travel times to work for 20 randomly selected New Yorkers

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

5	10	10	15	**15**	**15**	15	20	20	**20**	**25**	30	30	40	**40**	**45**	60	60	65	85

$Q_1 = 15$ $M = 22.5$ $Q_3 = 42.5$

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ &= 42.5 - 15 \\ &= 27.5 \text{ minutes} \end{aligned}$$

Interpretation: The range of the middle half of travel times for the New Yorkers in the sample is 27.5 minutes.

Identifying Outliers

- In addition to serving as a measure of spread, the interquartile range (IQR) is used as part of a rule of thumb for identifying outliers.

Definition:

The 1.5 x IQR Rule for Outliers

Call an observation an outlier if it falls more than 1.5 x IQR above the third quartile or below the first quartile.

Example, page 57

In the New York travel time data, we found Q_1=15 minutes, Q_3=42.5 minutes, and *IQR*=27.5 minutes.

For these data, 1.5 x *IQR* = 1.5(27.5) = 41.25

Q_1 - 1.5 x *IQR* = 15 – 41.25 = **-26.25**

Q_3+ 1.5 x *IQR* = 42.5 + 41.25 = **83.75**

Any travel time shorter than -26.25 minutes or longer than 83.75 minutes is considered an outlier.

0	5
1	005555
2	0005
3	00
4	005
5	
6	005
7	
8	**5**

The Five-Number Summary

- The minimum and maximum values alone tell us little about the distribution as a whole. Likewise, the median and quartiles tell us little about the tails of a distribution.
- To get a quick summary of both center and spread, combine all five numbers.

Definition:

The **five-number summary** of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest.

Minimum Q_1 ***M*** Q_3 ***Maximum***

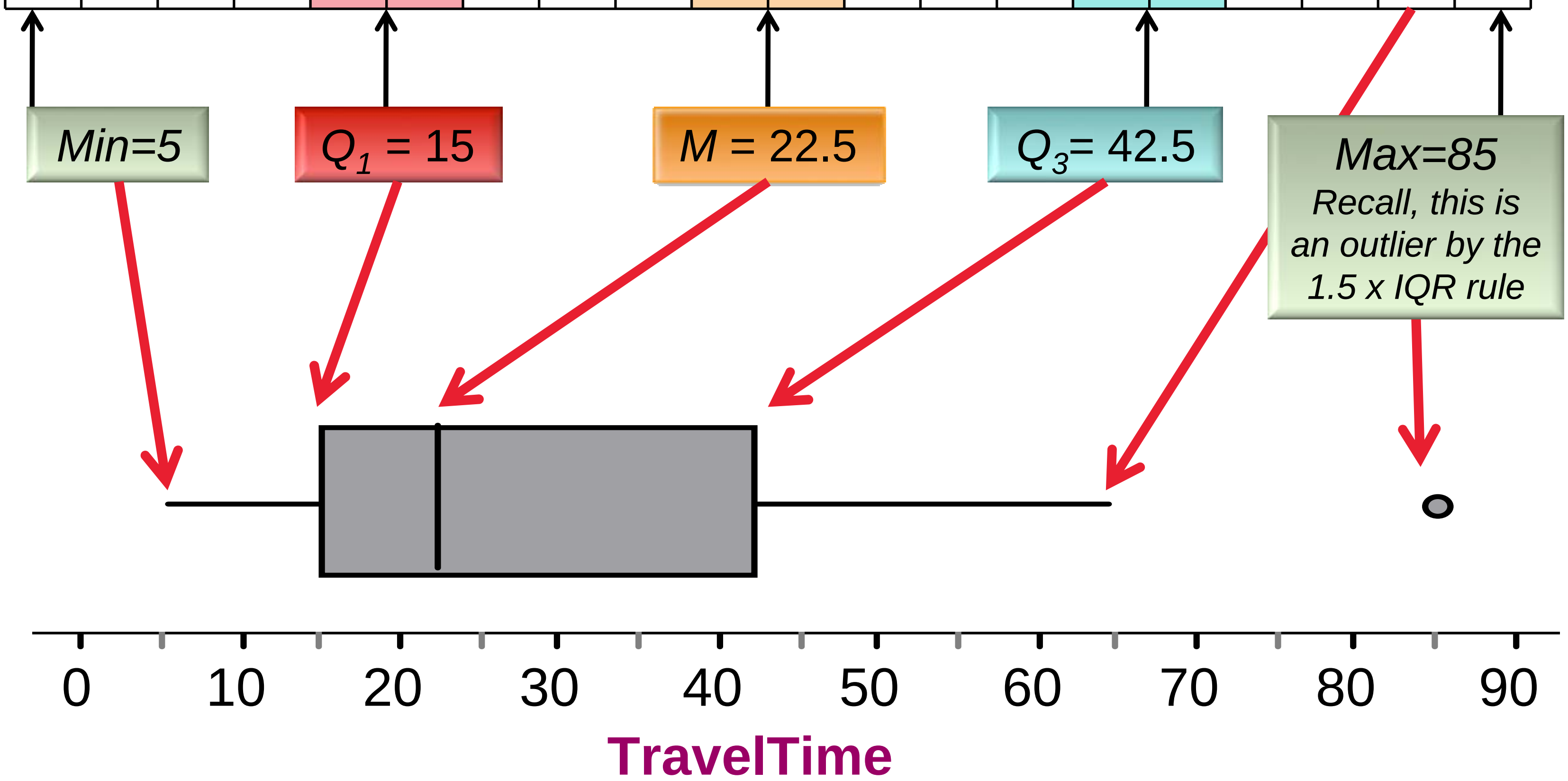
- The five-number summary divides the distribution roughly into quarters. This leads to a new way to display quantitative data, the **boxplot**.

Boxplots (Box-and-Whisker Plots)

- **Example: Consider our NY travel times data. Construct a boxplot.**
- **Note, Boxplots do not show the shape of our distribution. Use a histogram to see the shape.**

10	30	5	25	40	20	10	15	30	20	15	20	85	15	65	15	60	60	40	45

5	10	10	15	**15**	**15**	15	20	20	**20**	**25**	30	30	40	**40**	**45**	60	60	65	85



Measuring Spread: The Standard Deviation

Definition:

The **standard deviation** s_x measures the average distance of the observations from their mean.

It is calculated by finding an average of the squared distances and then taking the square root. This average squared distance is called the **variance**.

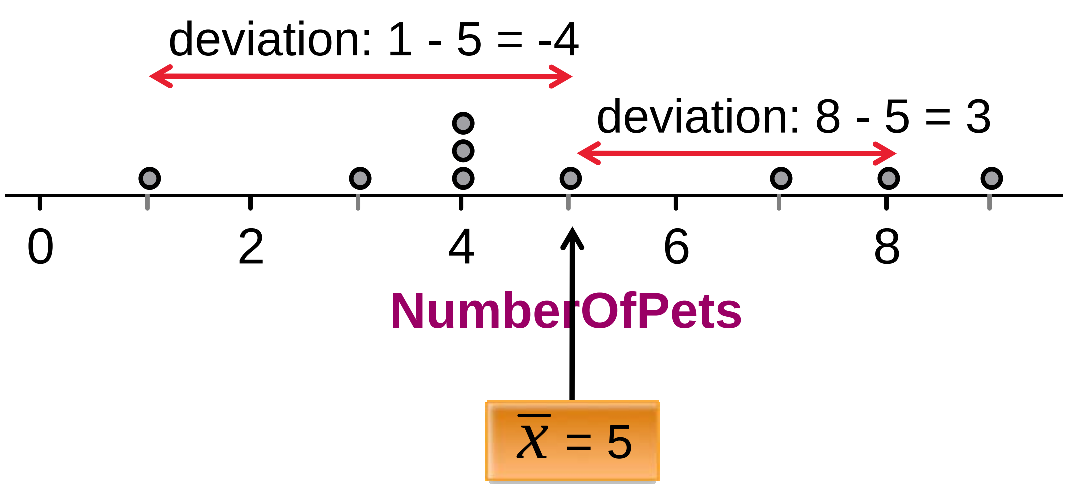
$$\text{variance} = s_x^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum (x_i - \bar{x})^2$$

$$\text{standard deviation} = s_x = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

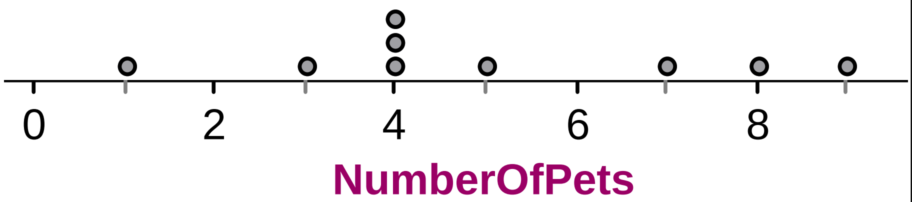
Measuring Spread: The Standard Deviation

- Let's explore it!
- Consider the following data on the number of pets owned by a group of 9 children.

1) Calculate the mean.
2) Calculate each *deviation.*
   deviation = observation – mean



Measuring Spread: The Standard Deviation



3) Square each deviation.

4) Find the "average" squared deviation. Calculate the sum of the squared deviations divided by (*n*-1)…this is called the **variance.**

5) Calculate the square root of the variance…this is the **standard deviation.**

x_i	$(x_i\text{-mean})$	$(x_i\text{-mean})^2$
1	1 - 5 = -4	$(-4)^2 = 16$
3	3 - 5 = -2	$(-2)^2 = 4$
4	4 - 5 = -1	$(-1)^2 = 1$
4	4 - 5 = -1	$(-1)^2 = 1$
4	4 - 5 = -1	$(-1)^2 = 1$
5	5 - 5 = 0	$(0)^2 = 0$
7	7 - 5 = 2	$(2)^2 = 4$
8	8 - 5 = 3	$(3)^2 = 9$
9	9 - 5 = 4	$(4)^2 = 16$
	Sum=0	**Sum=52**

"average" squared deviation = 52/(9-1) = 6.5 ← This is the **variance.**

Standard deviation = square root of variance = $\sqrt{6.5} = 2.55$

Resistant Measures

- We now have a choice between two descriptions for center and spread
  - Mean and Standard Deviation
  - Median and Interquartile Range

Choosing Measures of Center and Spread

•The median and *IQR* are usually better than the mean and standard deviation for describing a skewed distribution or a distribution with outliers.

•Use mean and standard deviation only for reasonably symmetric distributions that don't have outliers.

Introduction
Data Analysis: Making Sense of Data

- ✓ A **dataset** contains information on **individuals.**
- ✓ For each individual, data give values for one or more **variables.**
- ✓ Variables can be **categorical** or **quantitative.**
- ✓ The **distribution** of a variable describes what values it takes and how often it takes them.
- ✓ **Inference** is the process of making a conclusion about a population based on a sample set of data.

Summary

Section 1.1
Analyzing Categorical Data

- ✓ The distribution of a categorical variable lists the categories and gives the count or percent of individuals that fall into each category.
- ✓ **Pie charts** and **bar graphs** display the distribution of a categorical variable.
- ✓ A **two-way table** of counts organizes data about two categorical variables.
- ✓ The row-totals and column-totals in a two-way table give the **marginal distributions** of the two individual variables.
- ✓ There are two sets of **conditional distributions** for a two-way table.

Section 1.1
Analyzing Categorical Data

- ✓ We can use a **side-by-side bar graph** or a **segmented bar graph** to display conditional distributions.
- ✓ To describe the association between the row and column variables, compare an appropriate set of conditional distributions.
- ✓ Even a strong association between two categorical variables can be influenced by other variables lurking in the background.
- ✓ You can organize many problems using the four steps **state, plan, do,** and **conclude.**

Section 1.2

Displaying Quantitative Data with Graphs

- ✓ You can use a **dotplot, stemplot, or histogram** to show the distribution of a quantitative variable.
- ✓ When examining any graph, look for an **overall pattern** and for notable **departures** from that pattern. Describe the **shape, center, spread,** and any **outliers**. Don't forget your SOCS!
- ✓ Some distributions have simple shapes, such as **symmetric** or **skewed**. The number of **modes** (major peaks) is another aspect of overall shape.
- ✓ When comparing distributions, be sure to discuss shape, center, spread, and possible outliers.
- ✓ Histograms are for quantitative data, bar graphs are for categorical data. Use relative frequency histograms when comparing data sets of different sizes.

Section 1.3
Describing Quantitative Data with Numbers

- ✓ A numerical summary of a distribution should report at least its **center** and **spread**.
- ✓ The **mean** and **median** describe the center of a distribution in different ways. The mean is the average and the median is the midpoint of the values.
- ✓ When you use the median to indicate the center of a distribution, describe its spread using the **quartiles**.
- ✓ The **interquartile range (*IQR*)** is the range of the middle 50% of the observations: $IQR = Q_3 - Q_1$.

Section 1.3
Describing Quantitative Data with Numbers

Summary

- An extreme observation is an **outlier** if it is smaller than Q_1–(1.5x*IQR*) or larger than Q_3+(1.5x*IQR*) .
- The **five-number summary** (*min*, Q_1, *M*, Q_3, *max*) provides a quick overall description of distribution and can be pictured using a **boxplot.**
- The **variance** and its square root, the **standard deviation** are common measures of spread about the mean as center.
- The mean and standard deviation are good descriptions for symmetric distributions without outliers. The median and *IQR* are a better description for skewed distributions.

Organizing a Statistical Problem

- As you learn more about statistics, you will be asked to solve more complex problems.
- Here is a four-step process you can follow.

How to Organize a Statistical Problem: A Four-Step Process

State: What's the question that you're trying to answer?

Plan: How will you go about answering the question? What statistical techniques does this problem call for?

Do: Make graphs and carry out needed calculations.

Conclude: Give your practical conclusion in the setting of the real-world problem.