AP Statistics

Mr. Richard Denny, MBA

richard.denny@browardschools.com

Coral Glades High School

Room 215

What is AP Statistics?

AP Statistics is unlike any math class that you have taken. Every student always wants the answer to one question: "When will I ever use this?" The truth is of many other math disciplines is that while the theories and logic behind them are important, we rarely use complex math every day. Well, this is where AP Statistics differs. In this course, you will learn to use numbers presented in the data (statistics) to create accurate, research-based conclusions and predictions about the data.

An introductory **college-level** course, AP Statistics is an intensive look at the science of interpreting data. In this course, you will learn how to design, collect, organize, analyze and interpret data to create educated, accurate, research-based conclusions & predictions. The goal of this class is not to see a collection of numbers, but to see the meaning behind the numbers and fostering the ability to explain the data. A solid understanding of statistics will enable you to make a better analytical decision-maker in your career and everyday life.

Pre-Course Responsibilities

Prior to the beginning of the course, each student will need

- Complete Summer Work
- Invest in graphing calculator (preferably, TI-84+ Silver Edition or TI-84+ CE)
    - This makes calculations easier and less tedious and will help tremendously during the inferential statistics sections.

AP Statistics Summer Work

1. **Read Chapter 1 note slides or in the book (if available) and COMPLETE Chapter 1 note shell.** Answers to all questions can be found in the slides. Please read the slides **carefully.**
    - This **COMPLETED** note shell shall count as your first 3 homework assignments of the year.
    - Pay special attention to Highlighted items, **Emphasized wording** and **Alternately Colored Text** in the note shell.
    - Pay special attention to all boxes in the slides Labeled **"How to", "Properties" or "Caution"**

AP Statistics

2. **Complete Chapter 1 Review Exercises: #1-10 & the chapter 1 AP Statistics Practice Test at the end of the chapter**.
    - Please answer all questions using complete sentences to convey a clear and concise understanding of the information.
        - For example: If a question asks you to **calculate AND interpret**, then the math alone is not sufficient. It must be accompanied by a written explanation as to what that number means in the context of the question being asked. ***(Become VERY FAMILIAR with answering questions in this manner. It is the key to getting a passing score on the AP exam)***

3. Finally, please note there will be a **Chapter 1 Test** on our **3rd class meeting**. During the 1st and 2nd Class period, we will be reviewing chapter 1 key points. **PLEASE BE PREPARED!**

Topics that we cover in AP Statistics

Exploring Data (6 days)

The Normal Distributions (4 days)

Examining Data Relationships (4 days)

More on bivariate data (4 days)

Experimental Design: Producing Data (7 days)

Probability: The Study of Randomness (4 days)

Random Variables (3 days)

The Binomial and Geometric Distributions (4 days)

Sampling Distributions (6 days)

Introduction to Inference (5 days)

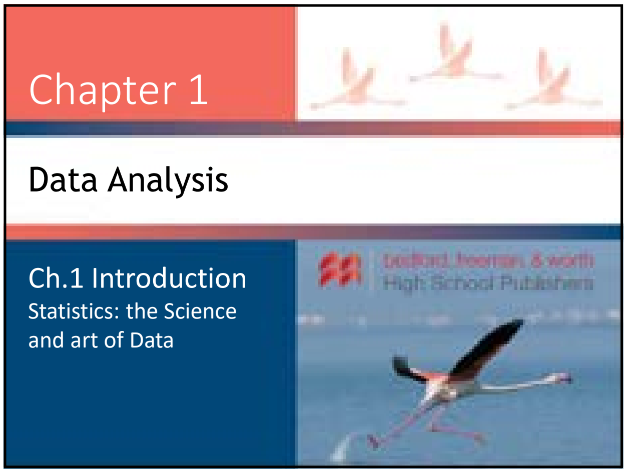
Inference for Distributions (5 days)

Inference for Proportions (3 days)

Inference for Tables: Chi-Square Procedures (4 days)

Inference for Regression (4 days)

Mandatory Final Project (Due after the AP exam): Counts as your Final Exam Grade



Chapter 1

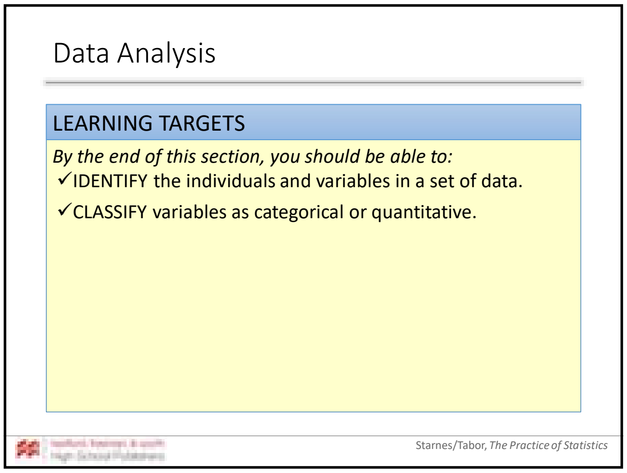
Data Analysis

Ch.1 Introduction

Statistics: the Science and art of Data

bedford, freeman & worth
High School Publishers

1



Data Analysis

LEARNING TARGETS

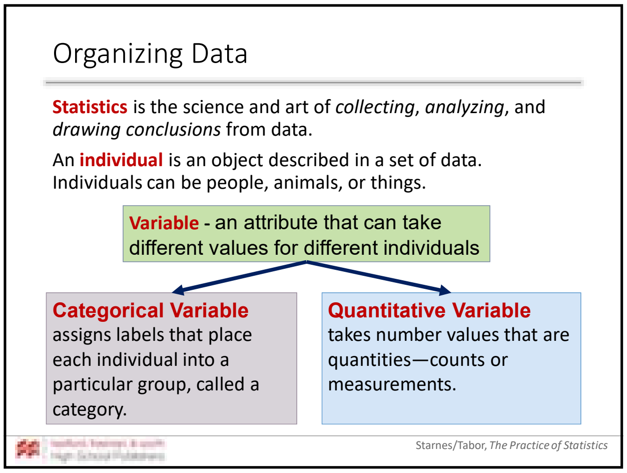
By the end of this section, you should be able to:

✓IDENTIFY the individuals and variables in a set of data.

✓CLASSIFY variables as categorical or quantitative.

Starnes/Tabor, *The Practice of Statistics*

2



Organizing Data

Statistics is the science and art of *collecting*, *analyzing*, and *drawing conclusions* from data.

An **individual** is an object described in a set of data. Individuals can be people, animals, or things.

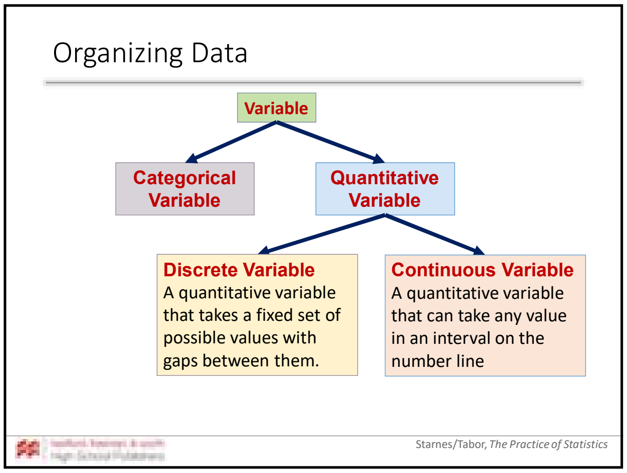
Variable - an attribute that can take different values for different individuals

Categorical Variable assigns labels that place each individual into a particular group, called a category.

Quantitative Variable takes number values that are quantities—counts or measurements.

Starnes/Tabor, *The Practice of Statistics*

3



Organizing Data

Variable

Categorical Variable

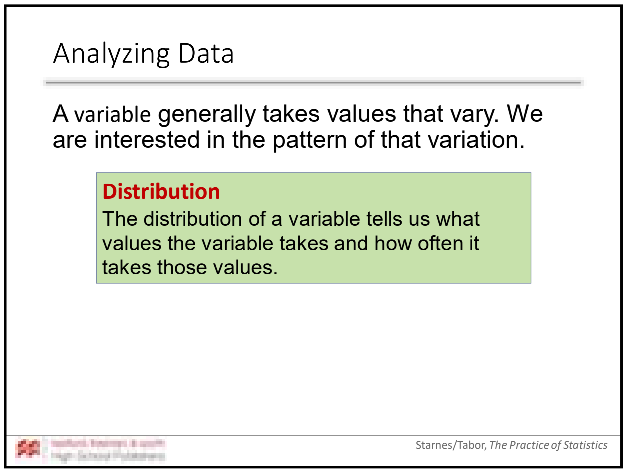
Quantitative Variable

Discrete Variable
A quantitative variable that takes a fixed set of possible values with gaps between them.

Continuous Variable
A quantitative variable that can take any value in an interval on the number line

Starnes/Tabor, *The Practice of Statistics*

4



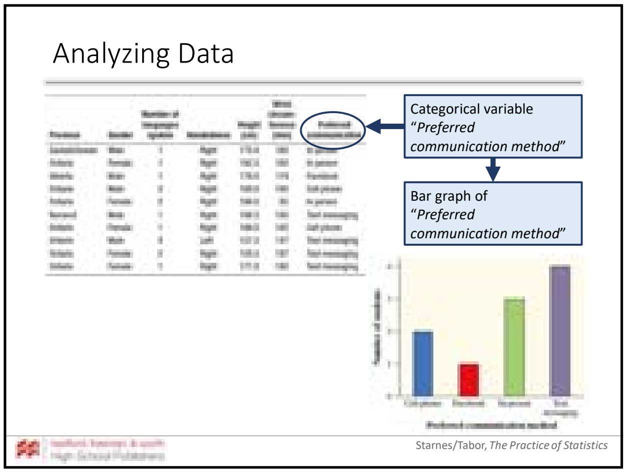
Analyzing Data

A variable generally takes values that vary. We are interested in the pattern of that variation.

Distribution
The distribution of a variable tells us what values the variable takes and how often it takes those values.

Starnes/Tabor, *The Practice of Statistics*

5



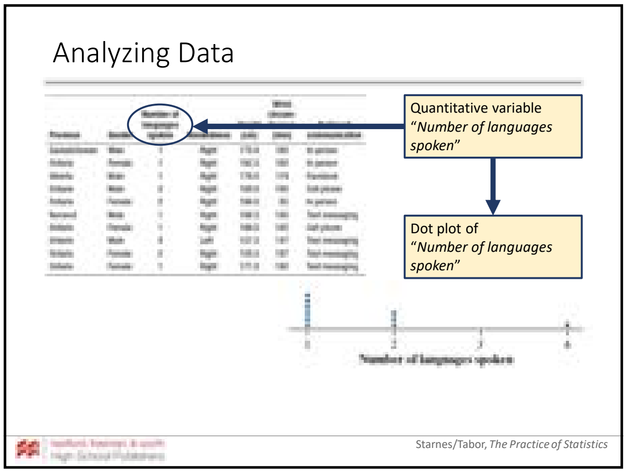
Analyzing Data

Categorical variable *“Preferred communication method”*

Bar graph of *“Preferred communication method”*

Starnes/Tabor, *The Practice of Statistics*

6



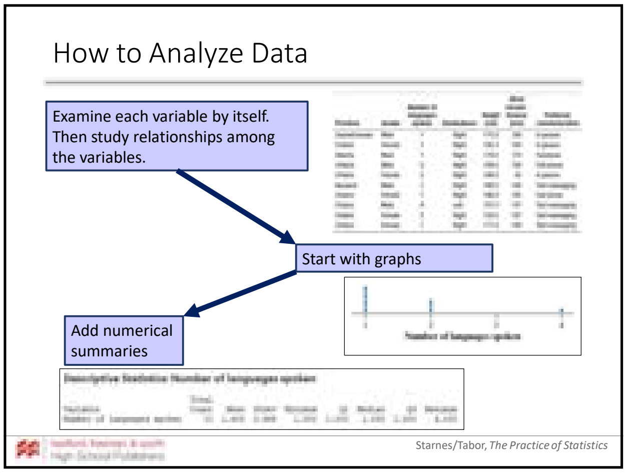
Analyzing Data

Quantitative variable *"Number of languages spoken"*

Dot plot of *"Number of languages spoken"*

Starnes/Tabor, *The Practice of Statistics*

7



How to Analyze Data

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

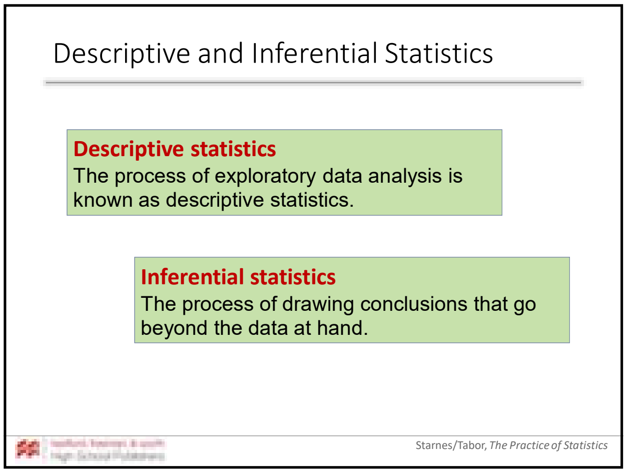
Start with graphs

Add numerical summaries

Number of languages spoken
Descriptive Statistics: Number of languages spoken

Starnes/Tabor, *The Practice of Statistics*

8



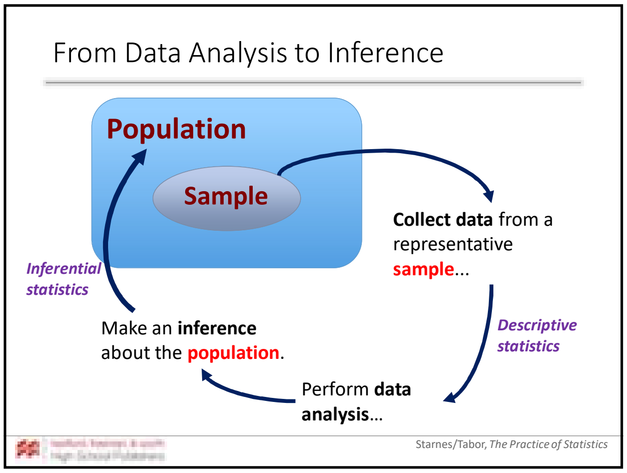
Descriptive and Inferential Statistics

Descriptive statistics
The process of exploratory data analysis is known as descriptive statistics.

Inferential statistics
The process of drawing conclusions that go beyond the data at hand.

Starnes/Tabor, *The Practice of Statistics*

9



From Data Analysis to Inference

Population

Sample

Collect data from a representative **sample**...

Descriptive statistics

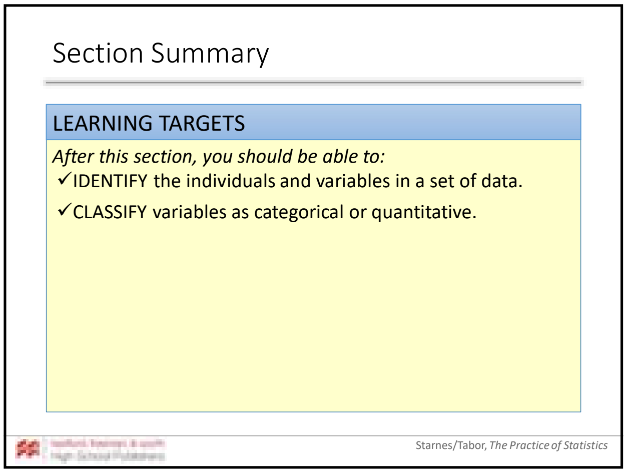
Perform **data analysis**...

Make an **inference** about the **population**.

Inferential statistics

Starnes/Tabor, *The Practice of Statistics*

10



Section Summary

LEARNING TARGETS

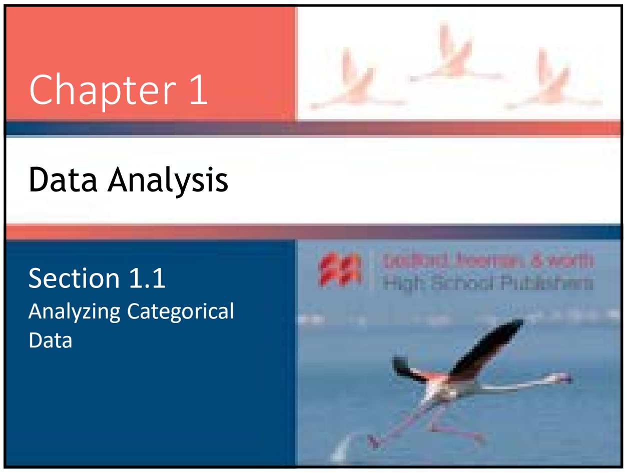
After this section, you should be able to:

✓IDENTIFY the individuals and variables in a set of data.

✓CLASSIFY variables as categorical or quantitative.

Starnes/Tabor, *The Practice of Statistics*

11



Chapter 1

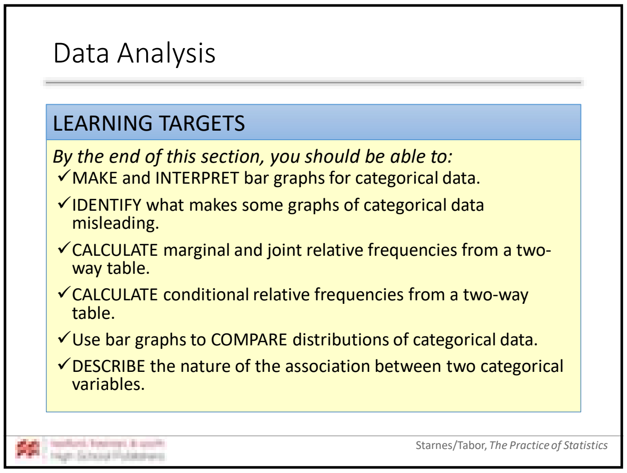
Data Analysis

Section 1.1

Analyzing Categorical Data

bedford, freeman & worth
High School Publishers

1



Data Analysis

LEARNING TARGETS

By the end of this section, you should be able to:

✓MAKE and INTERPRET bar graphs for categorical data.

✓IDENTIFY what makes some graphs of categorical data misleading.

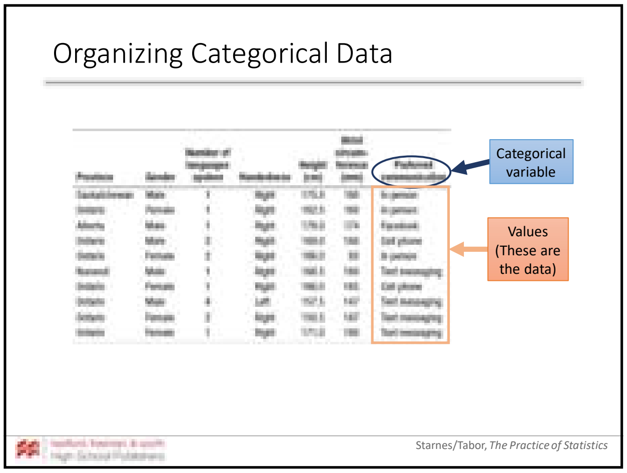
✓CALCULATE marginal and joint relative frequencies from a two-way table.

✓CALCULATE conditional relative frequencies from a two-way table.

✓Use bar graphs to COMPARE distributions of categorical data.

✓DESCRIBE the nature of the association between two categorical variables.

2



Organizing Categorical Data

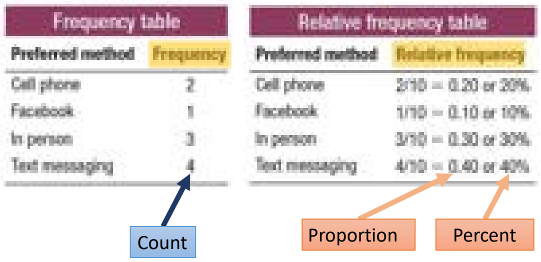
Province	Gender	Number of languages spoken	Handedness	Height (cm)	Wrist circumference (mm)	Preferred communication
Saskatchewan	Male	1	Right	175.0	180	In person
Ontario	Female	1	Right	162.5	160	In person
Alberta	Male	1	Right	178.0	174	Facebook
Ontario	Male	2	Right	169.0	160	Cell phone
Ontario	Female	2	Right	166.0	65	In person
Nunavut	Male	1	Right	168.5	160	Text messaging
Ontario	Female	1	Right	166.0	165	Cell phone
Ontario	Male	4	Left	157.5	147	Text messaging
Ontario	Female	2	Right	150.5	187	Text messaging
Ontario	Female	1	Right	171.0	180	Text messaging

Categorical variable

Values (These are the data)

3

Organizing Categorical Data



Frequency table	
Preferred method	Frequency
Cell phone	2
Facebook	1
In person	3
Text messaging	4

Relative frequency table	
Preferred method	Relative frequency
Cell phone	2/10 = 0.20 or 20%
Facebook	1/10 = 0.10 or 10%
In person	3/10 = 0.30 or 30%
Text messaging	4/10 = 0.40 or 40%

Count

Proportion

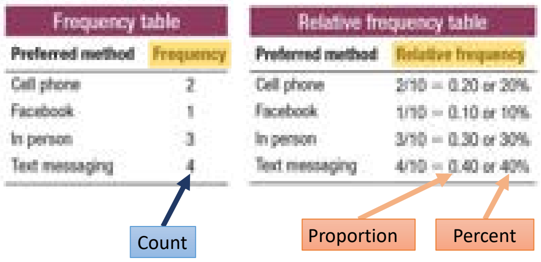
Percent

Starnes/Tabor, *The Practice of Statistics*

4

Displaying Categorical Data

To display the distribution of categorical data, make a **bar graph**



Frequency table	
Preferred method	Frequency
Cell phone	2
Facebook	1
In person	3
Text messaging	4

Relative frequency table	
Preferred method	Relative frequency
Cell phone	2/10 = 0.20 or 20%
Facebook	1/10 = 0.10 or 10%
In person	3/10 = 0.30 or 30%
Text messaging	4/10 = 0.40 or 40%

Count

Proportion

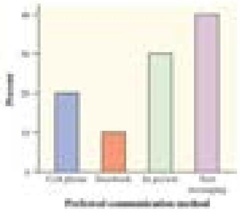
Percent

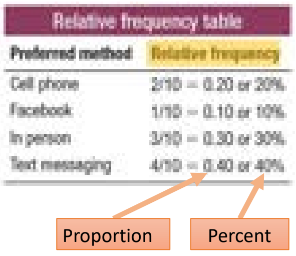
Starnes/Tabor, *The Practice of Statistics*

5

Displaying Categorical Data

To display the distribution of categorical data, make a **bar graph**





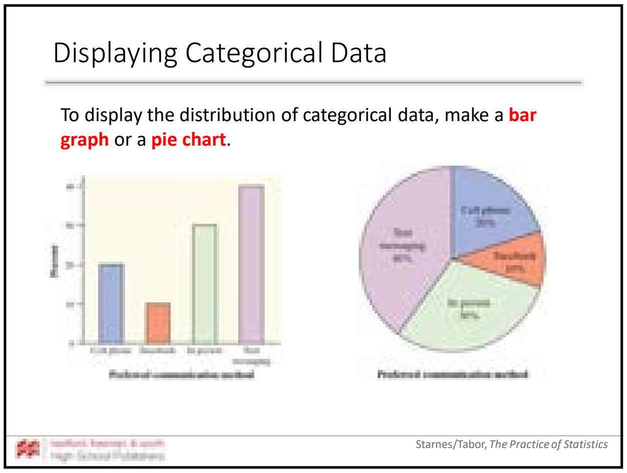
Relative frequency table	
Preferred method	Relative frequency
Cell phone	2/10 = 0.20 or 20%
Facebook	1/10 = 0.10 or 10%
In person	3/10 = 0.30 or 30%
Text messaging	4/10 = 0.40 or 40%

Proportion

Percent

Starnes/Tabor, *The Practice of Statistics*

6

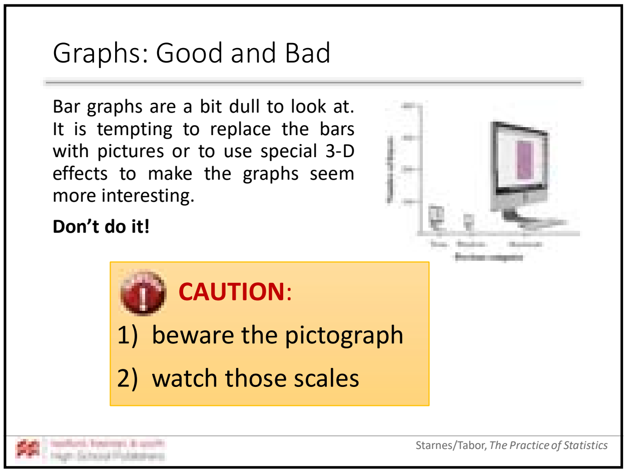


Displaying Categorical Data

To display the distribution of categorical data, make a **bar graph** or a **pie chart**.

Starnes/Tabor, *The Practice of Statistics*

7



Graphs: Good and Bad

Bar graphs are a bit dull to look at. It is tempting to replace the bars with pictures or to use special 3-D effects to make the graphs seem more interesting.

Don't do it!

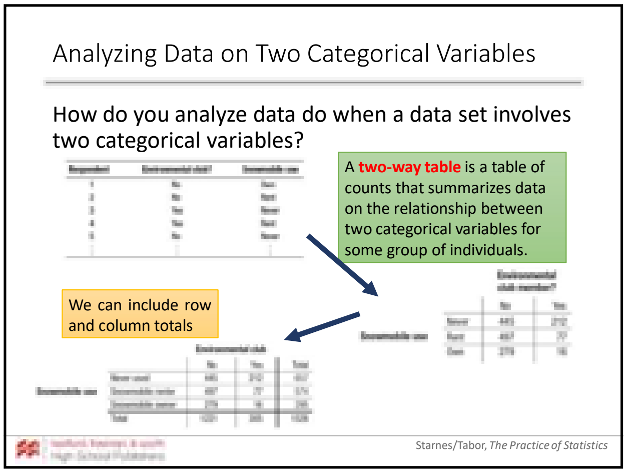
CAUTION:

1) beware the pictograph

2) watch those scales

Starnes/Tabor, *The Practice of Statistics*

8



Analyzing Data on Two Categorical Variables

How do you analyze data do when a data set involves two categorical variables?

Respondent	Environmental club?	Snowmobile use
1	No	Own
2	No	Rent
3	Yes	Never
4	Yes	Rent
5	No	Never
⋮	⋮	⋮

A **two-way table** is a table of counts that summarizes data on the relationship between two categorical variables for some group of individuals.

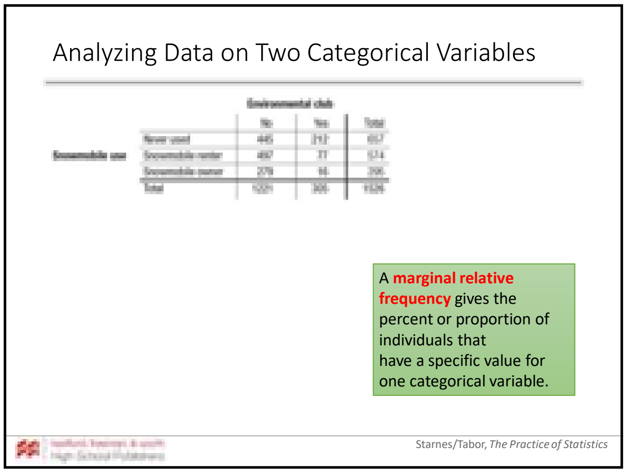
		Environmental club member?	
		No	Yes
Snowmobile use	Never	445	212
	Rent	497	77
	Own	279	16

We can include row and column totals

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

Starnes/Tabor, *The Practice of Statistics*

9



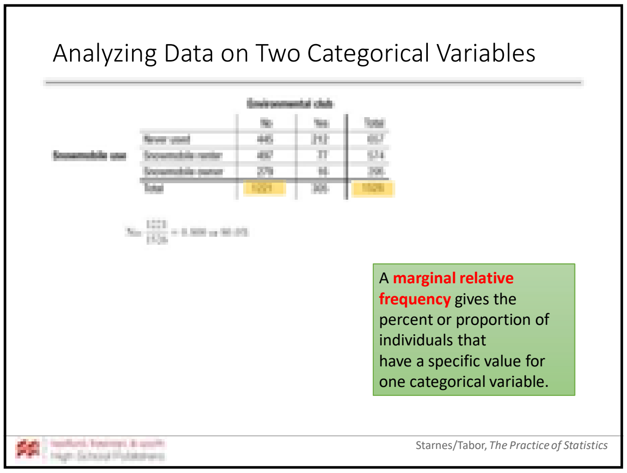
Analyzing Data on Two Categorical Variables

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

A **marginal relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable.

Starnes/Tabor, *The Practice of Statistics*

10



Analyzing Data on Two Categorical Variables

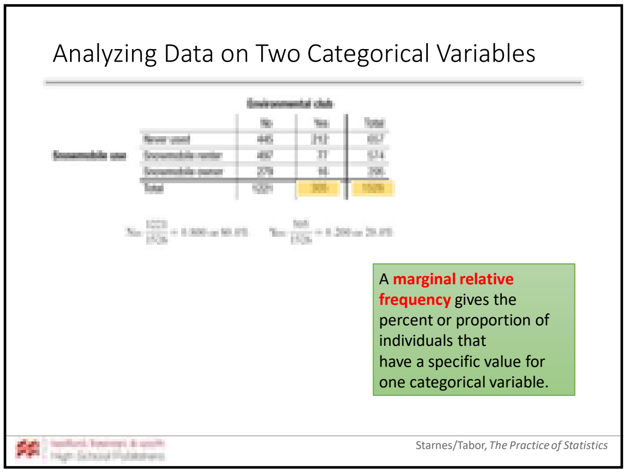
		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

No: $\frac{1221}{1526} = 0.800$ or 80.0%

A **marginal relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable.

Starnes/Tabor, *The Practice of Statistics*

11



Analyzing Data on Two Categorical Variables

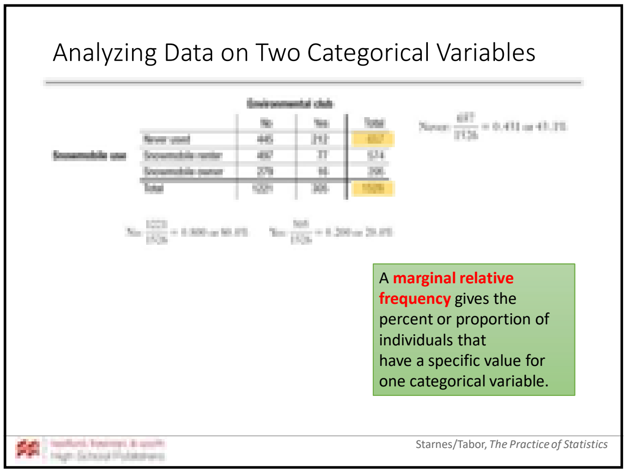
		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

No: $\frac{1221}{1526} = 0.800$ or 80.0% Yes: $\frac{305}{1526} = 0.200$ or 20.0%

A **marginal relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable.

Starnes/Tabor, *The Practice of Statistics*

12



Analyzing Data on Two Categorical Variables

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

Never: $\frac{657}{1526} = 0.431$ or 43.1%

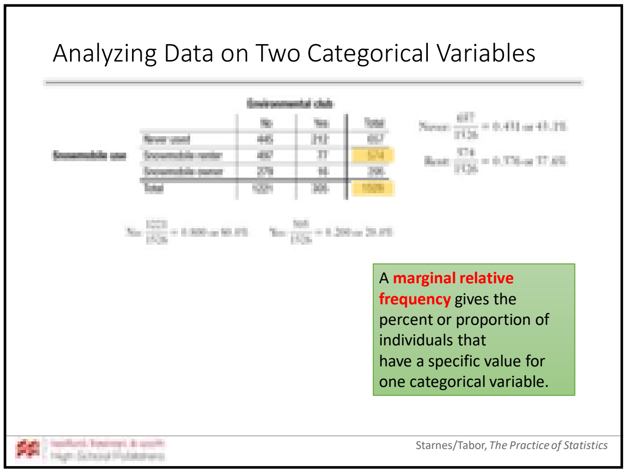
No: $\frac{1221}{1526} = 0.800$ or 80.0% Yes: $\frac{305}{1526} = 0.200$ or 20.0%

A **marginal relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable.

Bedford, Freeman & Worth High School Publishers

Starnes/Tabor, *The Practice of Statistics*

13



Analyzing Data on Two Categorical Variables

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

Never: $\frac{657}{1526} = 0.431$ or 43.1%

Rent: $\frac{574}{1526} = 0.376$ or 37.6%

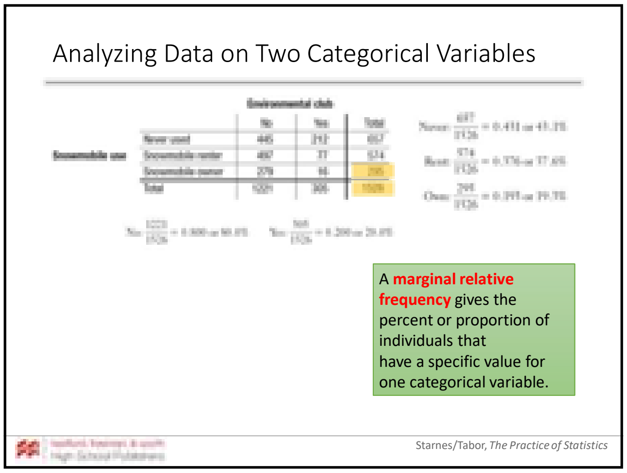
No: $\frac{1221}{1526} = 0.800$ or 80.0% Yes: $\frac{305}{1526} = 0.200$ or 20.0%

A **marginal relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable.

Bedford, Freeman & Worth High School Publishers

Starnes/Tabor, *The Practice of Statistics*

14



Analyzing Data on Two Categorical Variables

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

Never: $\frac{657}{1526} = 0.431$ or 43.1%

Rent: $\frac{574}{1526} = 0.376$ or 37.6%

Own: $\frac{295}{1526} = 0.193$ or 19.3%

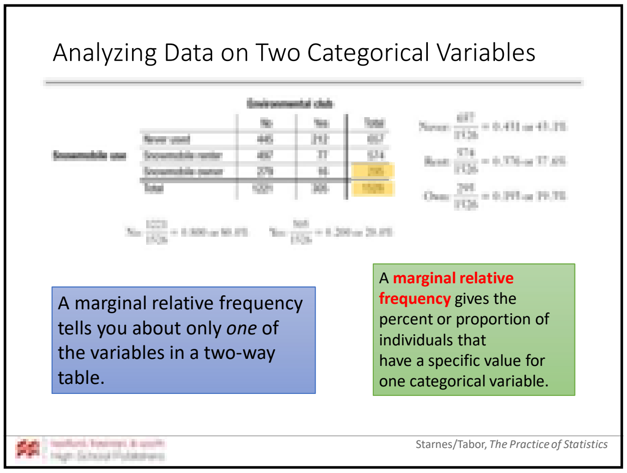
No: $\frac{1221}{1526} = 0.800$ or 80.0% Yes: $\frac{305}{1526} = 0.200$ or 20.0%

A **marginal relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable.

Bedford, Freeman & Worth High School Publishers

Starnes/Tabor, *The Practice of Statistics*

15



Analyzing Data on Two Categorical Variables

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

Never: $\frac{657}{1526} = 0.431$ or 43.1%

Rent: $\frac{574}{1526} = 0.376$ or 37.6%

Own: $\frac{295}{1526} = 0.193$ or 19.3%

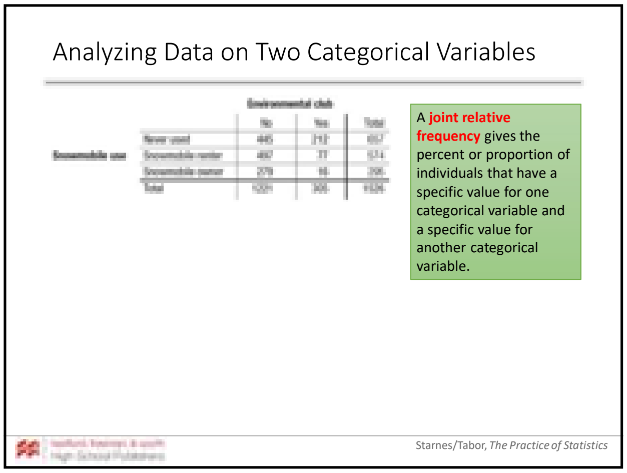
No: $\frac{1221}{1526} = 0.800$ or 80.0%   Yes: $\frac{305}{1526} = 0.200$ or 20.0%

A marginal relative frequency tells you about only *one* of the variables in a two-way table.

A **marginal relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable.

Starnes/Tabor, *The Practice of Statistics*

16



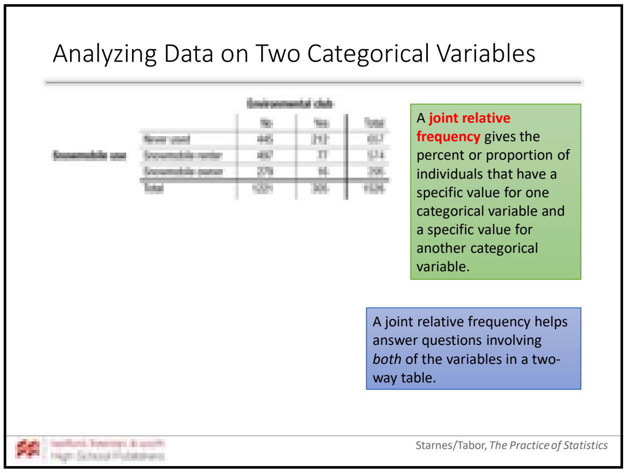
Analyzing Data on Two Categorical Variables

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

A **joint relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable and a specific value for another categorical variable.

Starnes/Tabor, *The Practice of Statistics*

17



Analyzing Data on Two Categorical Variables

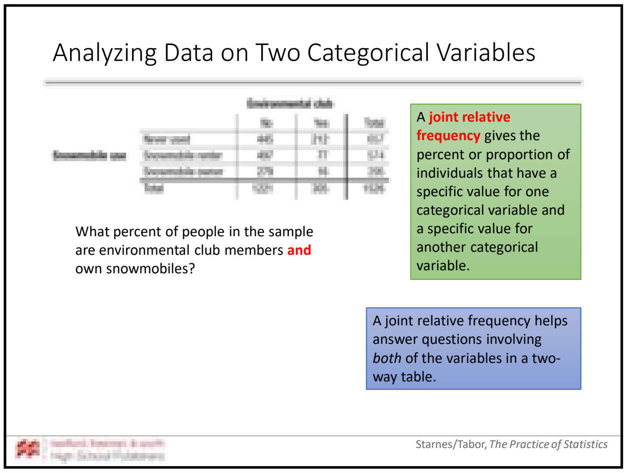
		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

A **joint relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable and a specific value for another categorical variable.

A joint relative frequency helps answer questions involving *both* of the variables in a two-way table.

Starnes/Tabor, *The Practice of Statistics*

18



Analyzing Data on Two Categorical Variables

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

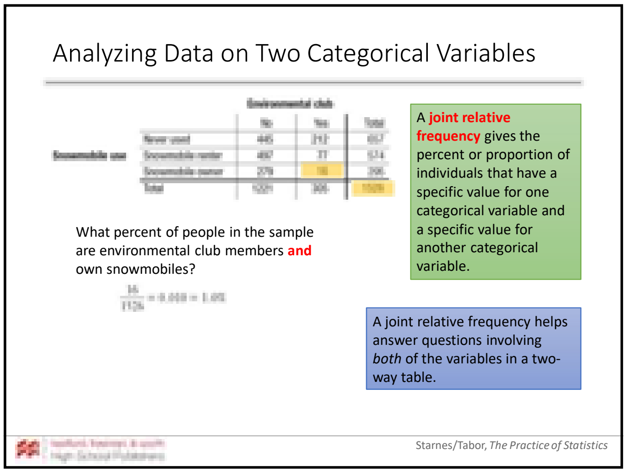
What percent of people in the sample are environmental club members **and** own snowmobiles?

A **joint relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable and a specific value for another categorical variable.

A joint relative frequency helps answer questions involving *both* of the variables in a two-way table.

Starnes/Tabor, *The Practice of Statistics*

19



Analyzing Data on Two Categorical Variables

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

What percent of people in the sample are environmental club members **and** own snowmobiles?

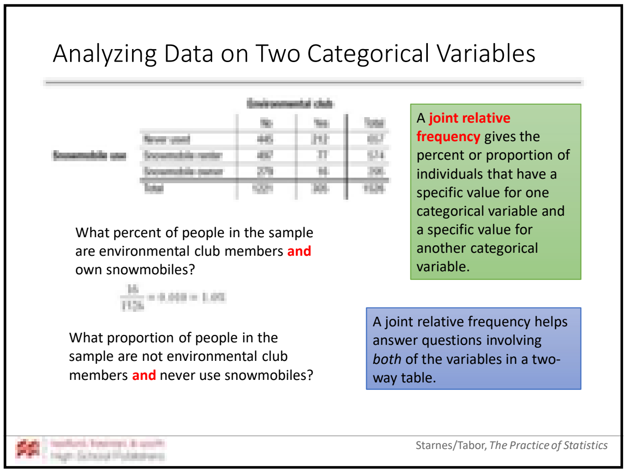
$\frac{16}{1526} = 0.010 = 1.0\%$

A **joint relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable and a specific value for another categorical variable.

A joint relative frequency helps answer questions involving *both* of the variables in a two-way table.

Starnes/Tabor, *The Practice of Statistics*

20



Analyzing Data on Two Categorical Variables

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

What percent of people in the sample are environmental club members **and** own snowmobiles?

$\frac{16}{1526} = 0.010 = 1.0\%$

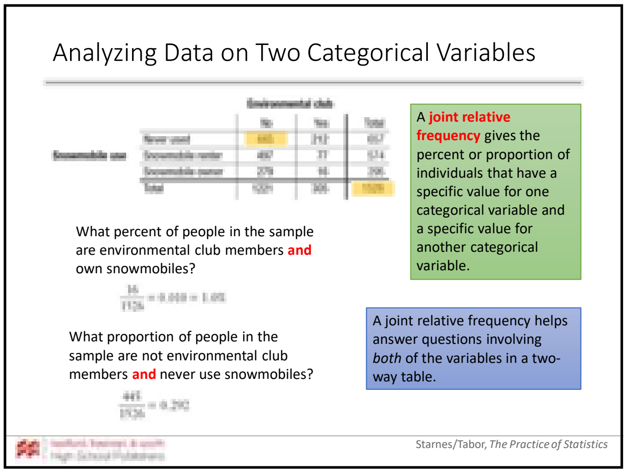
What proportion of people in the sample are not environmental club members **and** never use snowmobiles?

A **joint relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable and a specific value for another categorical variable.

A joint relative frequency helps answer questions involving *both* of the variables in a two-way table.

Starnes/Tabor, *The Practice of Statistics*

21



Analyzing Data on Two Categorical Variables

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

A **joint relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable and a specific value for another categorical variable.

What percent of people in the sample are environmental club members **and** own snowmobiles?

$$\frac{16}{1526} = 0.010 = 1.0\%$$

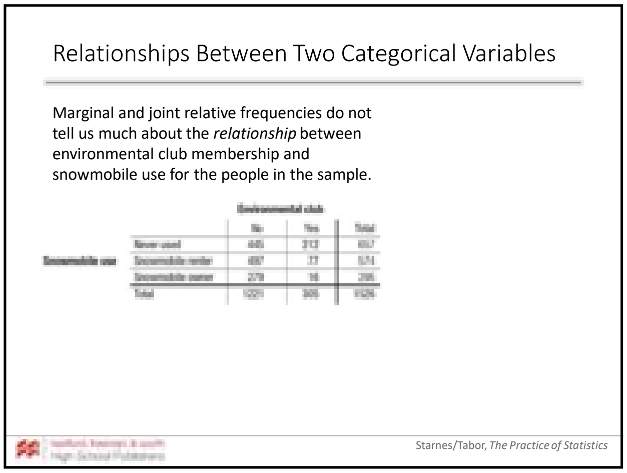
What proportion of people in the sample are not environmental club members **and** never use snowmobiles?

$$\frac{445}{1526} = 0.292$$

A joint relative frequency helps answer questions involving *both* of the variables in a two-way table.

Starnes/Tabor, *The Practice of Statistics*

22



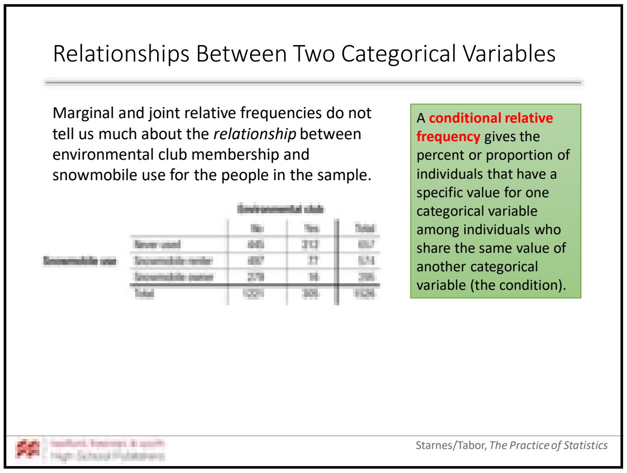
Relationships Between Two Categorical Variables

Marginal and joint relative frequencies do not tell us much about the *relationship* between environmental club membership and snowmobile use for the people in the sample.

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

Starnes/Tabor, *The Practice of Statistics*

23



Relationships Between Two Categorical Variables

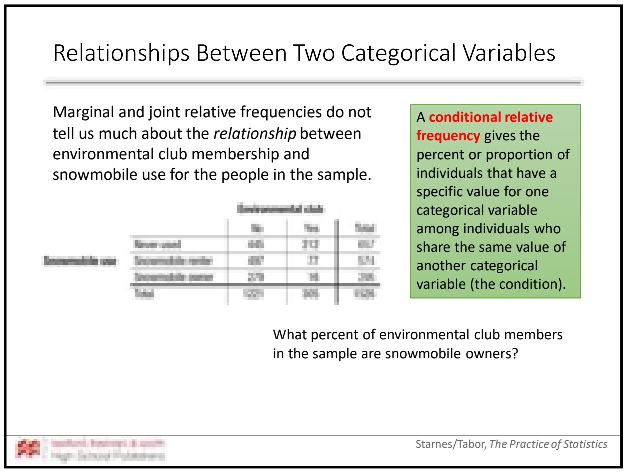
Marginal and joint relative frequencies do not tell us much about the *relationship* between environmental club membership and snowmobile use for the people in the sample.

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

A **conditional relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable among individuals who share the same value of another categorical variable (the condition).

Starnes/Tabor, *The Practice of Statistics*

24



Relationships Between Two Categorical Variables

Marginal and joint relative frequencies do not tell us much about the *relationship* between environmental club membership and snowmobile use for the people in the sample.

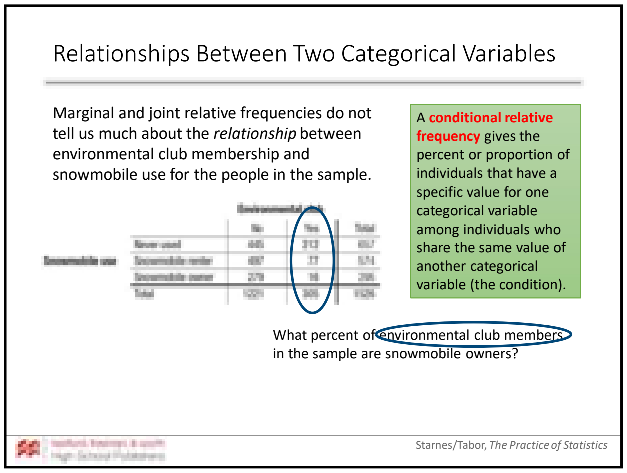
		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

A **conditional relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable among individuals who share the same value of another categorical variable (the condition).

What percent of environmental club members in the sample are snowmobile owners?

Starnes/Tabor, *The Practice of Statistics*

25



Relationships Between Two Categorical Variables

Marginal and joint relative frequencies do not tell us much about the *relationship* between environmental club membership and snowmobile use for the people in the sample.

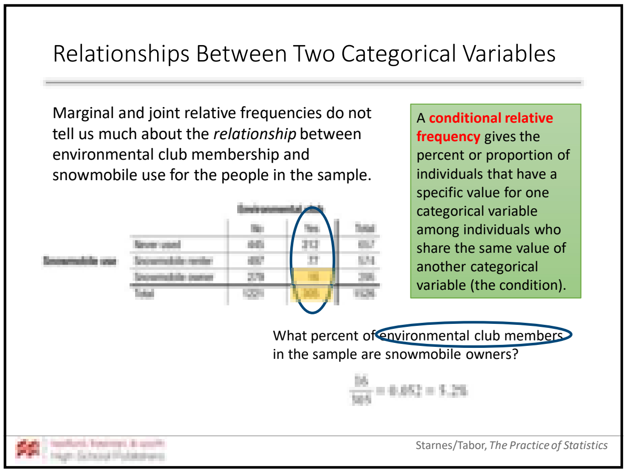
		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

A **conditional relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable among individuals who share the same value of another categorical variable (the condition).

What percent of environmental club members in the sample are snowmobile owners?

Starnes/Tabor, *The Practice of Statistics*

26



Relationships Between Two Categorical Variables

Marginal and joint relative frequencies do not tell us much about the *relationship* between environmental club membership and snowmobile use for the people in the sample.

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

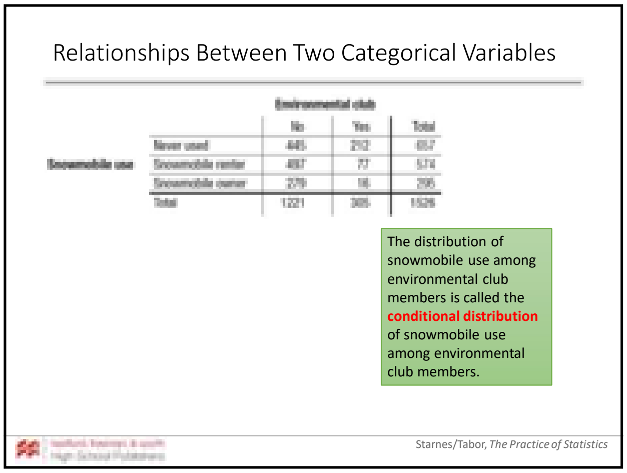
A **conditional relative frequency** gives the percent or proportion of individuals that have a specific value for one categorical variable among individuals who share the same value of another categorical variable (the condition).

What percent of environmental club members in the sample are snowmobile owners?

$$\frac{16}{305} = 0.052 = 5.2\%$$

Starnes/Tabor, *The Practice of Statistics*

27



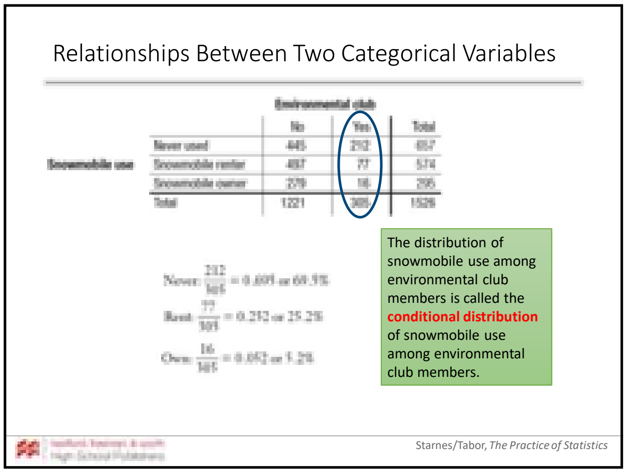
Relationships Between Two Categorical Variables

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

The distribution of snowmobile use among environmental club members is called the **conditional distribution** of snowmobile use among environmental club members.

Starnes/Tabor, *The Practice of Statistics*

28



Relationships Between Two Categorical Variables

		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

Never: $\frac{212}{305} = 0.695$ or 69.5%

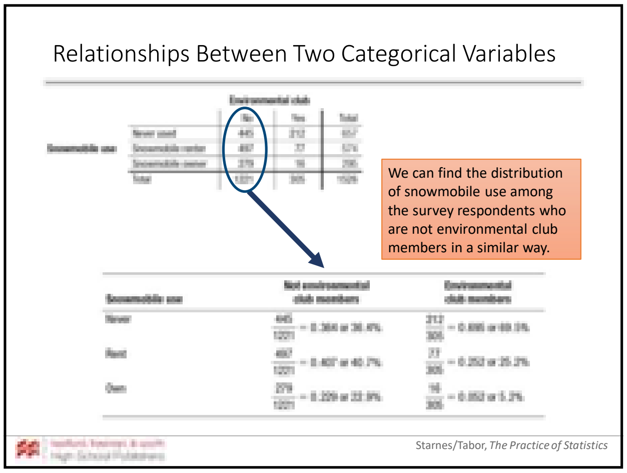
Rent: $\frac{77}{305} = 0.252$ or 25.2%

Own: $\frac{16}{305} = 0.052$ or 5.2%

The distribution of snowmobile use among environmental club members is called the **conditional distribution** of snowmobile use among environmental club members.

Starnes/Tabor, *The Practice of Statistics*

29



Relationships Between Two Categorical Variables

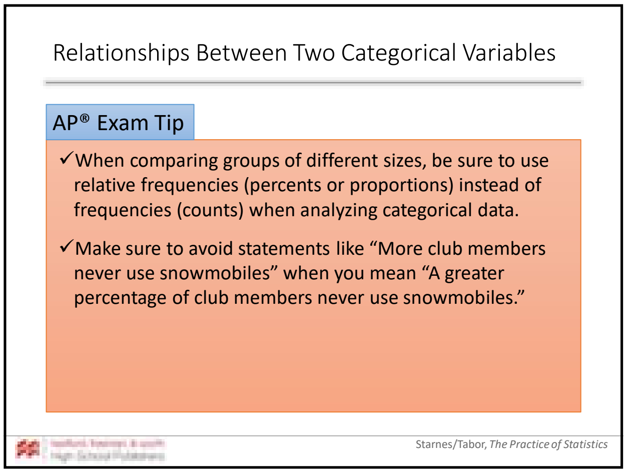
		Environmental club		
		No	Yes	Total
Snowmobile use	Never used	445	212	657
	Snowmobile renter	497	77	574
	Snowmobile owner	279	16	295
	Total	1221	305	1526

We can find the distribution of snowmobile use among the survey respondents who are not environmental club members in a similar way.

Snowmobile use	Not environmental club members	Environmental club members
Never	$\frac{445}{1221} = 0.364$ or 36.4%	$\frac{212}{305} = 0.695$ or 69.5%
Rent	$\frac{497}{1221} = 0.407$ or 40.7%	$\frac{77}{305} = 0.252$ or 25.2%
Own	$\frac{279}{1221} = 0.229$ or 22.9%	$\frac{16}{305} = 0.052$ or 5.2%

Starnes/Tabor, *The Practice of Statistics*

30



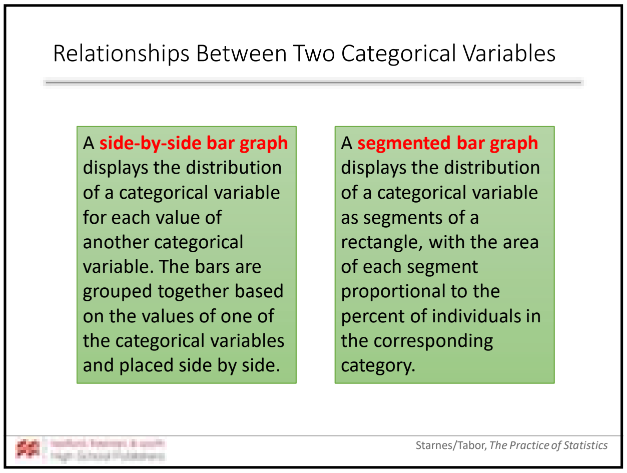
Relationships Between Two Categorical Variables

AP® Exam Tip

- ✓When comparing groups of different sizes, be sure to use relative frequencies (percents or proportions) instead of frequencies (counts) when analyzing categorical data.
- ✓Make sure to avoid statements like “More club members never use snowmobiles” when you mean “A greater percentage of club members never use snowmobiles.”

Starnes/Tabor, *The Practice of Statistics*

31



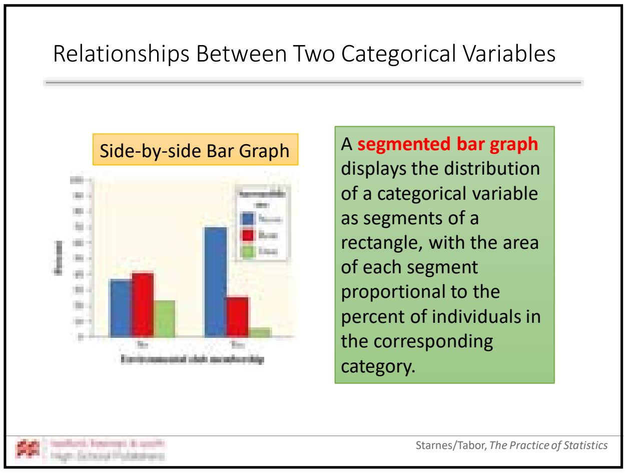
Relationships Between Two Categorical Variables

A **side-by-side bar graph** displays the distribution of a categorical variable for each value of another categorical variable. The bars are grouped together based on the values of one of the categorical variables and placed side by side.

A **segmented bar graph** displays the distribution of a categorical variable as segments of a rectangle, with the area of each segment proportional to the percent of individuals in the corresponding category.

Starnes/Tabor, *The Practice of Statistics*

32



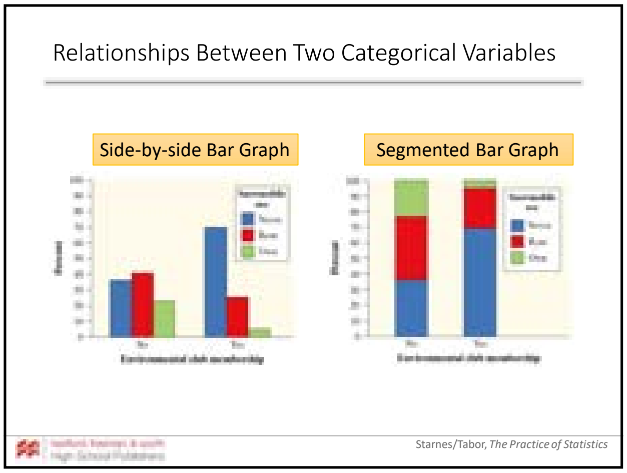
Relationships Between Two Categorical Variables

Side-by-side Bar Graph

A **segmented bar graph** displays the distribution of a categorical variable as segments of a rectangle, with the area of each segment proportional to the percent of individuals in the corresponding category.

Starnes/Tabor, *The Practice of Statistics*

33



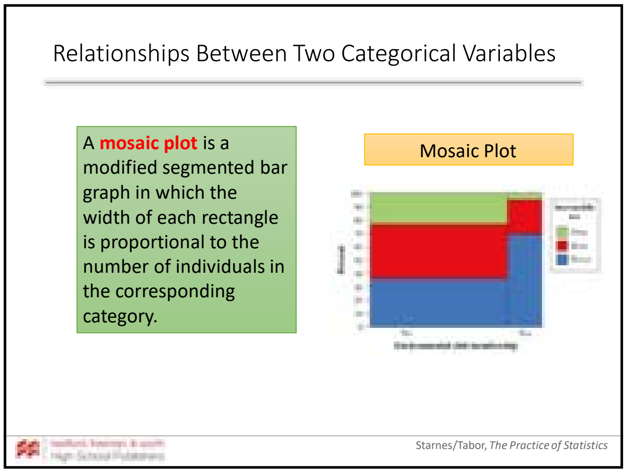
Relationships Between Two Categorical Variables

Side-by-side Bar Graph

Segmented Bar Graph

Starnes/Tabor, *The Practice of Statistics*

34



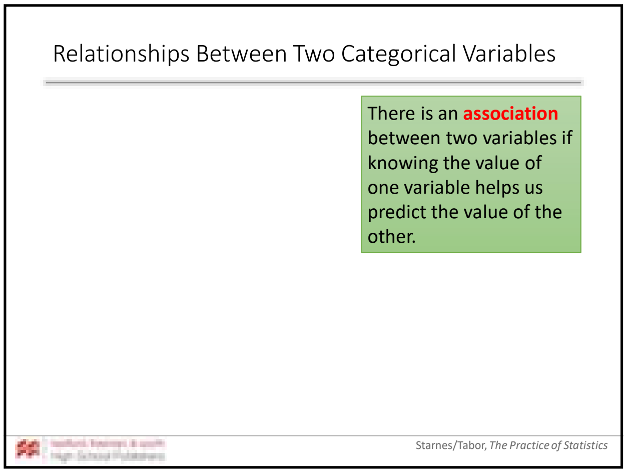
Relationships Between Two Categorical Variables

A **mosaic plot** is a modified segmented bar graph in which the width of each rectangle is proportional to the number of individuals in the corresponding category.

Mosaic Plot

Starnes/Tabor, *The Practice of Statistics*

35

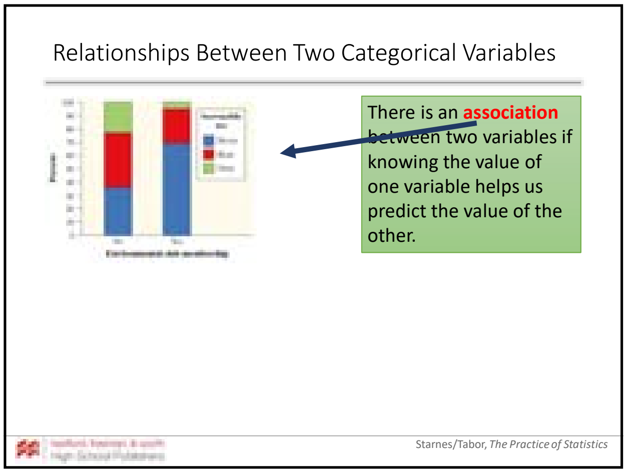


Relationships Between Two Categorical Variables

There is an **association** between two variables if knowing the value of one variable helps us predict the value of the other.

Starnes/Tabor, *The Practice of Statistics*

36

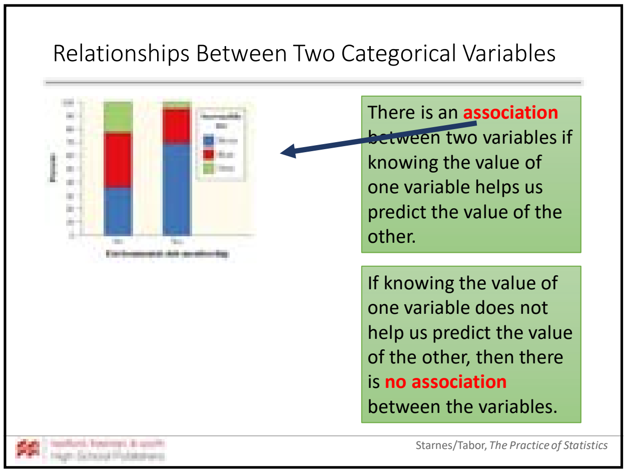


Relationships Between Two Categorical Variables

There is an **association** between two variables if knowing the value of one variable helps us predict the value of the other.

Starnes/Tabor, *The Practice of Statistics*

37



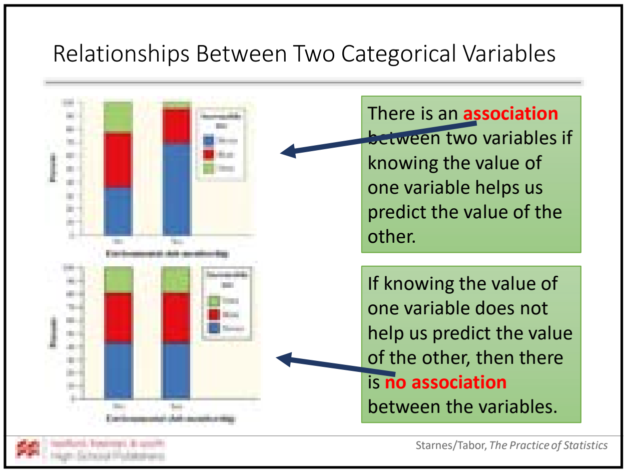
Relationships Between Two Categorical Variables

There is an **association** between two variables if knowing the value of one variable helps us predict the value of the other.

If knowing the value of one variable does not help us predict the value of the other, then there is **no association** between the variables.

Starnes/Tabor, *The Practice of Statistics*

38



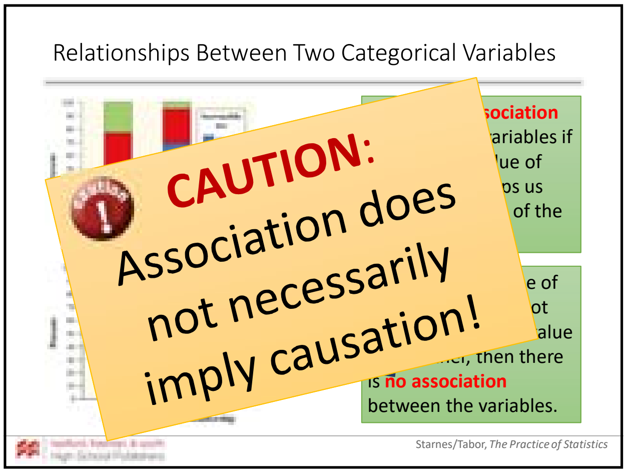
Relationships Between Two Categorical Variables

There is an **association** between two variables if knowing the value of one variable helps us predict the value of the other.

If knowing the value of one variable does not help us predict the value of the other, then there is **no association** between the variables.

Starnes/Tabor, *The Practice of Statistics*

39



Relationships Between Two Categorical Variables

CAUTION: Association does not necessarily imply causation!

sociation
ariables if
lue of
ps us
of the

e of
ot
alue
ner, then there
is **no association**
between the variables.

Starnes/Tabor, *The Practice of Statistics*

40

Section Summary

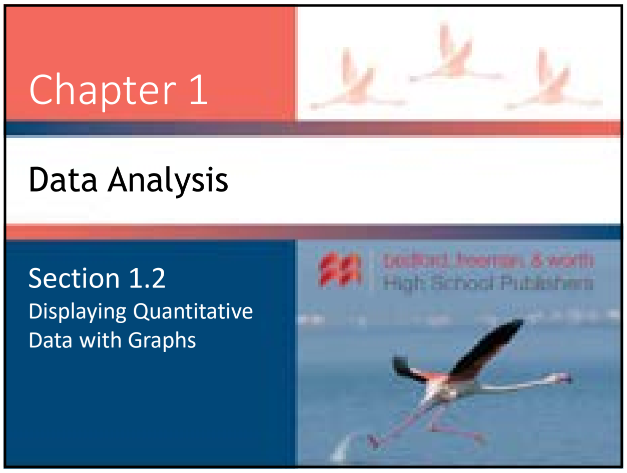
LEARNING TARGETS

After this section, you should be able to:

- ✓MAKE and INTERPRET bar graphs for categorical data.
- ✓IDENTIFY what makes some graphs of categorical data misleading.
- ✓CALCULATE marginal and joint relative frequencies from a two-way table.
- ✓CALCULATE conditional relative frequencies from a two-way table.
- ✓Use bar graphs to COMPARE distributions of categorical data.
- ✓DESCRIBE the nature of the association between two categorical variables.

Starnes/Tabor, *The Practice of Statistics*

41



Chapter 1

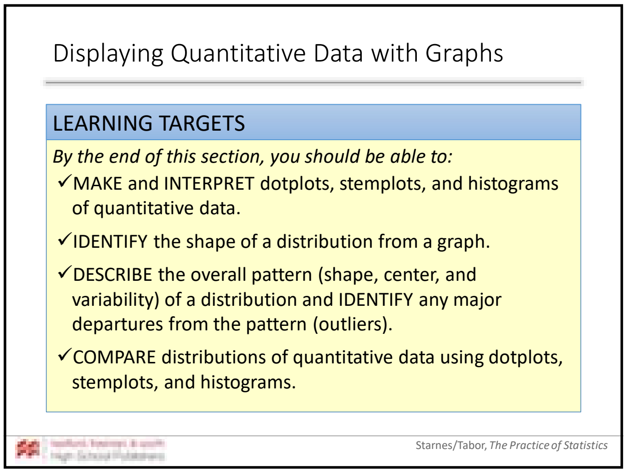
Data Analysis

Section 1.2

Displaying Quantitative Data with Graphs

bedford, freeman & worth
High School Publishers

1



Displaying Quantitative Data with Graphs

LEARNING TARGETS

By the end of this section, you should be able to:

✓MAKE and INTERPRET dotplots, stemplots, and histograms of quantitative data.

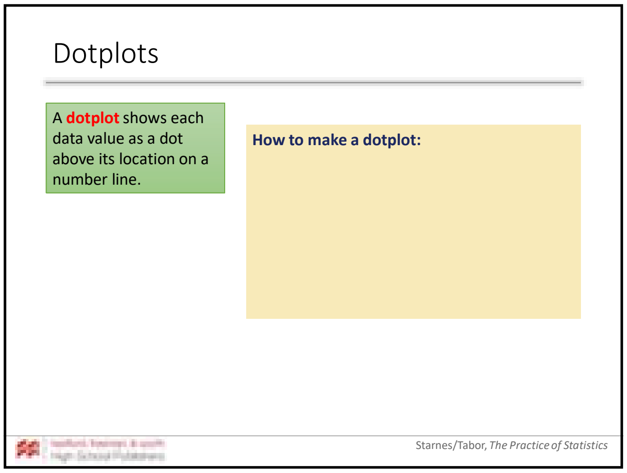
✓IDENTIFY the shape of a distribution from a graph.

✓DESCRIBE the overall pattern (shape, center, and variability) of a distribution and IDENTIFY any major departures from the pattern (outliers).

✓COMPARE distributions of quantitative data using dotplots, stemplots, and histograms.

Starnes/Tabor, *The Practice of Statistics*

2



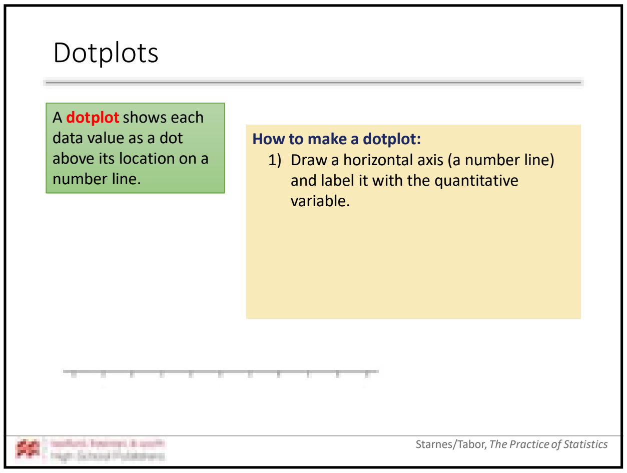
Dotplots

A **dotplot** shows each data value as a dot above its location on a number line.

How to make a dotplot:

Starnes/Tabor, *The Practice of Statistics*

3



Dotplots

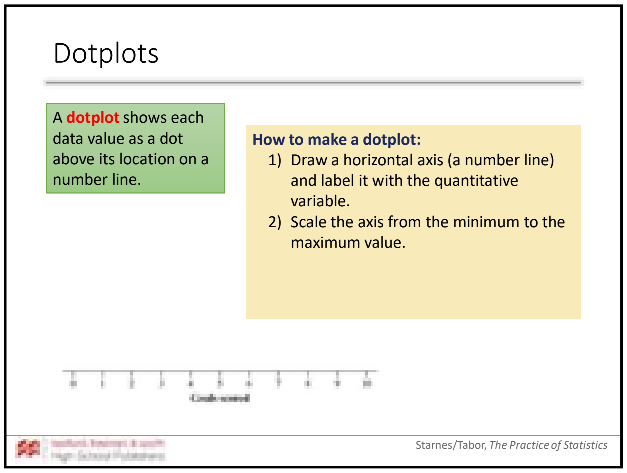
A **dotplot** shows each data value as a dot above its location on a number line.

How to make a dotplot:

1) Draw a horizontal axis (a number line) and label it with the quantitative variable.

Starnes/Tabor, *The Practice of Statistics*

4



Dotplots

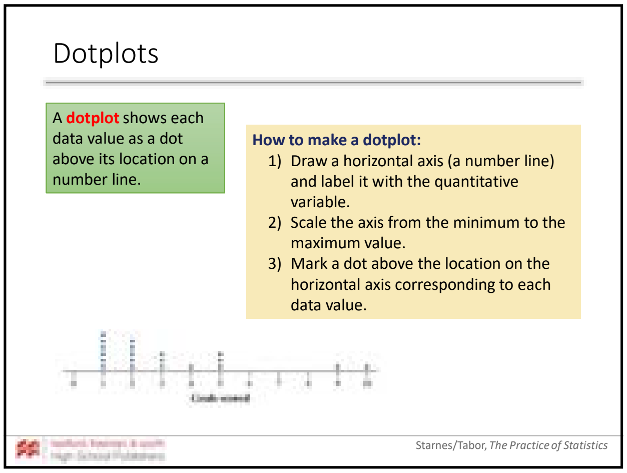
A **dotplot** shows each data value as a dot above its location on a number line.

How to make a dotplot:

1) Draw a horizontal axis (a number line) and label it with the quantitative variable.
2) Scale the axis from the minimum to the maximum value.

Starnes/Tabor, *The Practice of Statistics*

5



Dotplots

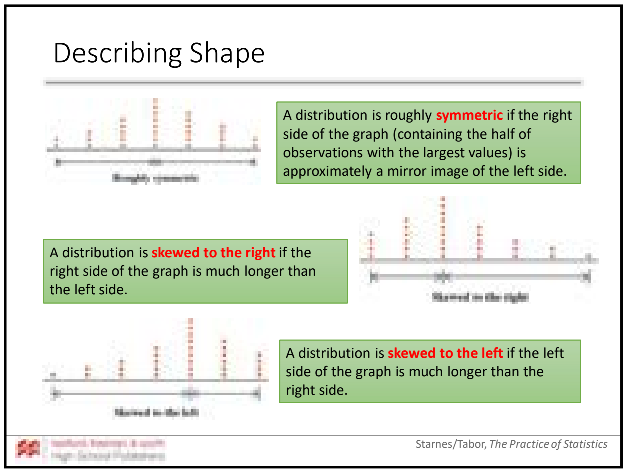
A **dotplot** shows each data value as a dot above its location on a number line.

How to make a dotplot:

1) Draw a horizontal axis (a number line) and label it with the quantitative variable.
2) Scale the axis from the minimum to the maximum value.
3) Mark a dot above the location on the horizontal axis corresponding to each data value.

Starnes/Tabor, *The Practice of Statistics*

6



Describing Shape

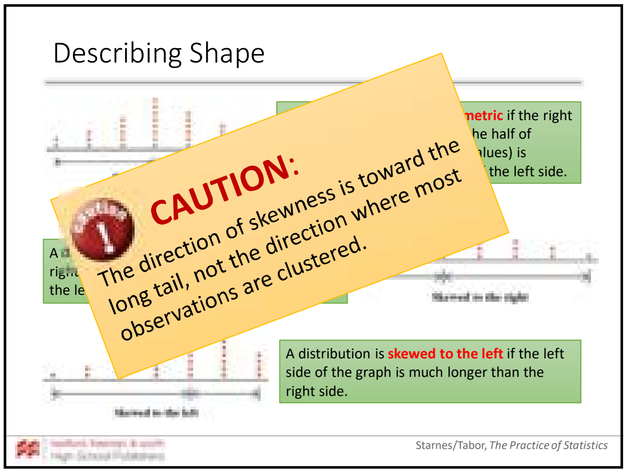
A distribution is roughly **symmetric** if the right side of the graph (containing the half of observations with the largest values) is approximately a mirror image of the left side.

A distribution is **skewed to the right** if the right side of the graph is much longer than the left side.

A distribution is **skewed to the left** if the left side of the graph is much longer than the right side.

Starnes/Tabor, *The Practice of Statistics*

7



Describing Shape

CAUTION: The direction of skewness is toward the long tail, not the direction where most observations are clustered.

metric if the right he half of alues) is the left side.

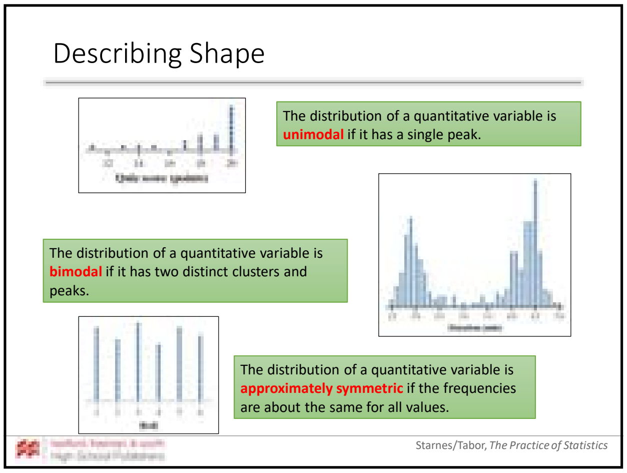
A d right the le

A distribution is **skewed to the left** if the left side of the graph is much longer than the right side.

Skewed to the right
Skewed to the left

Starnes/Tabor, *The Practice of Statistics*

8



Describing Shape

The distribution of a quantitative variable is **unimodal** if it has a single peak.

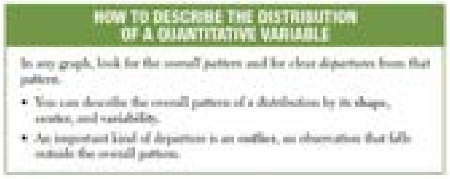
The distribution of a quantitative variable is **bimodal** if it has two distinct clusters and peaks.

The distribution of a quantitative variable is **approximately symmetric** if the frequencies are about the same for all values.

Starnes/Tabor, *The Practice of Statistics*

9

Describing Distributions



HOW TO DESCRIBE THE DISTRIBUTION OF A QUANTITATIVE VARIABLE

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

- You can describe the overall pattern of a distribution by its shape, center, and variability.
- An important kind of departure is an outlier, an observation that falls outside the overall pattern.

AP® Exam Tip

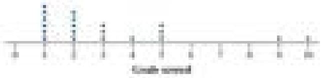
Always be sure to include context when you are asked to describe a distribution. This means using the variable name, not just the units the variable is measured in.

10

Describing Distributions

Describe the distribution of goals scored in 20 games played by the 2016 U.S. women's soccer team.

Did we include context?



Shape: The distribution of goals scored is skewed to the right, with a single peak at 1 goal. There is a gap between 5 and 9 goals.
Outliers: The games when the team scored 9 and 10 goals appear to be outliers.
Center: The median is 2 goals scored.
Variability: The number of goals varies from 1 to 10 goals scored.

11

Describing Distributions

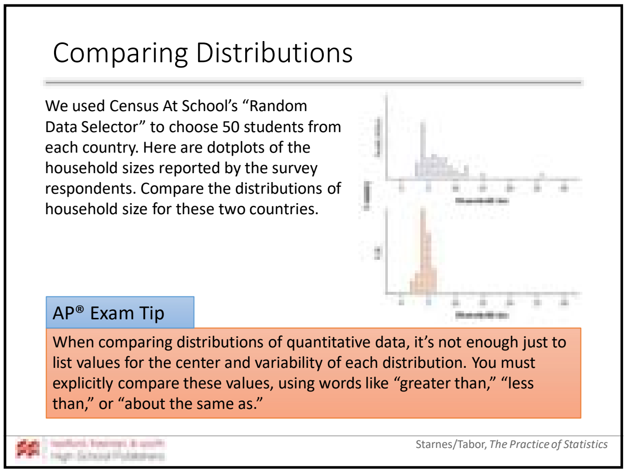
Describe the distribution of goals scored in 20 games played by the 2016 U.S. women's soccer team.

Did we include context? **YES!**

0
1
2
3
4
5
6
7
8
9
10
Goals scored

Shape: The distribution of goals scored is skewed to the right, with a single peak at 1 goal. There is a gap between 5 and 9 goals.
Outliers: The games when the team scored 9 and 10 goals appear to be outliers.
Center: The median is 2 goals scored.
Variability: The number of goals varies from 1 to 10 goals scored.

12



Comparing Distributions

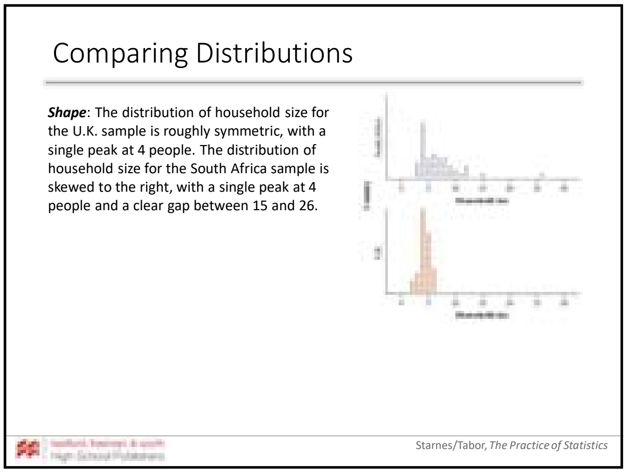
We used Census At School's "Random Data Selector" to choose 50 students from each country. Here are dotplots of the household sizes reported by the survey respondents. Compare the distributions of household size for these two countries.

AP® Exam Tip

When comparing distributions of quantitative data, it's not enough just to list values for the center and variability of each distribution. You must explicitly compare these values, using words like "greater than," "less than," or "about the same as."

Starnes/Tabor, *The Practice of Statistics*

13

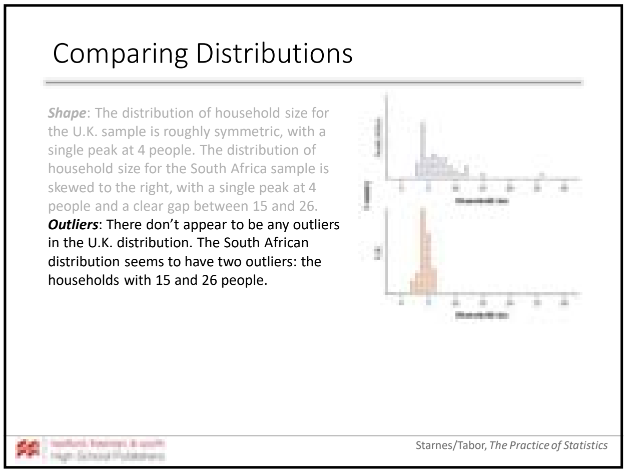


Comparing Distributions

Shape: The distribution of household size for the U.K. sample is roughly symmetric, with a single peak at 4 people. The distribution of household size for the South Africa sample is skewed to the right, with a single peak at 4 people and a clear gap between 15 and 26.

Starnes/Tabor, *The Practice of Statistics*

14



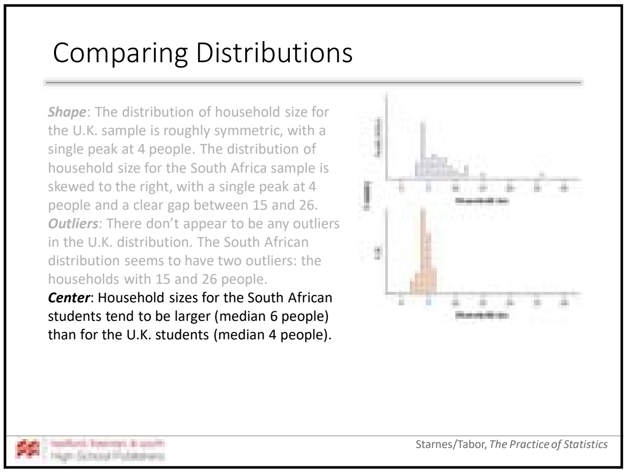
Comparing Distributions

Shape: The distribution of household size for the U.K. sample is roughly symmetric, with a single peak at 4 people. The distribution of household size for the South Africa sample is skewed to the right, with a single peak at 4 people and a clear gap between 15 and 26.

Outliers: There don't appear to be any outliers in the U.K. distribution. The South African distribution seems to have two outliers: the households with 15 and 26 people.

Starnes/Tabor, *The Practice of Statistics*

15

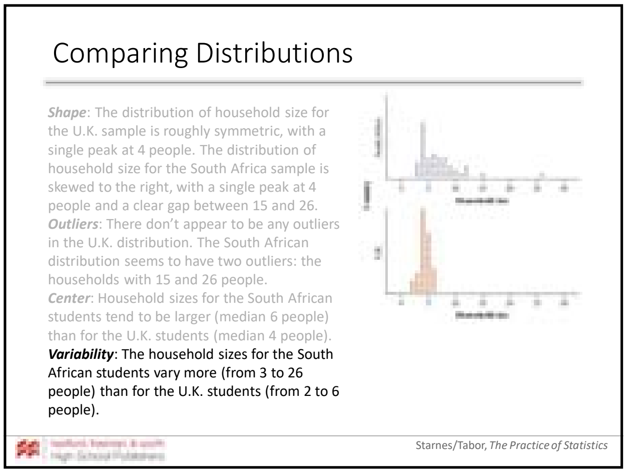


Comparing Distributions

Shape: The distribution of household size for the U.K. sample is roughly symmetric, with a single peak at 4 people. The distribution of household size for the South Africa sample is skewed to the right, with a single peak at 4 people and a clear gap between 15 and 26.
Outliers: There don't appear to be any outliers in the U.K. distribution. The South African distribution seems to have two outliers: the households with 15 and 26 people.
Center: Household sizes for the South African students tend to be larger (median 6 people) than for the U.K. students (median 4 people).

Starnes/Tabor, *The Practice of Statistics*

16

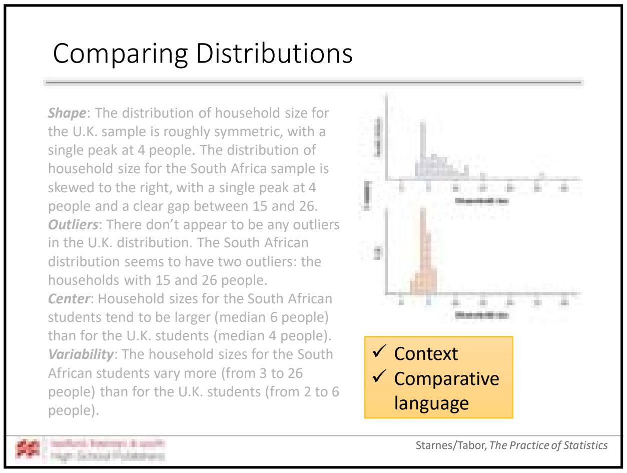


Comparing Distributions

Shape: The distribution of household size for the U.K. sample is roughly symmetric, with a single peak at 4 people. The distribution of household size for the South Africa sample is skewed to the right, with a single peak at 4 people and a clear gap between 15 and 26.
Outliers: There don't appear to be any outliers in the U.K. distribution. The South African distribution seems to have two outliers: the households with 15 and 26 people.
Center: Household sizes for the South African students tend to be larger (median 6 people) than for the U.K. students (median 4 people).
Variability: The household sizes for the South African students vary more (from 3 to 26 people) than for the U.K. students (from 2 to 6 people).

Starnes/Tabor, *The Practice of Statistics*

17



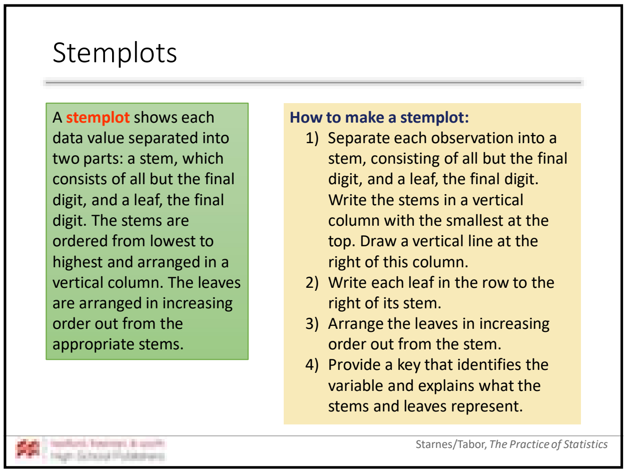
Comparing Distributions

Shape: The distribution of household size for the U.K. sample is roughly symmetric, with a single peak at 4 people. The distribution of household size for the South Africa sample is skewed to the right, with a single peak at 4 people and a clear gap between 15 and 26.
Outliers: There don't appear to be any outliers in the U.K. distribution. The South African distribution seems to have two outliers: the households with 15 and 26 people.
Center: Household sizes for the South African students tend to be larger (median 6 people) than for the U.K. students (median 4 people).
Variability: The household sizes for the South African students vary more (from 3 to 26 people) than for the U.K. students (from 2 to 6 people).

✓ Context
✓ Comparative language

Starnes/Tabor, *The Practice of Statistics*

18



Stemplots

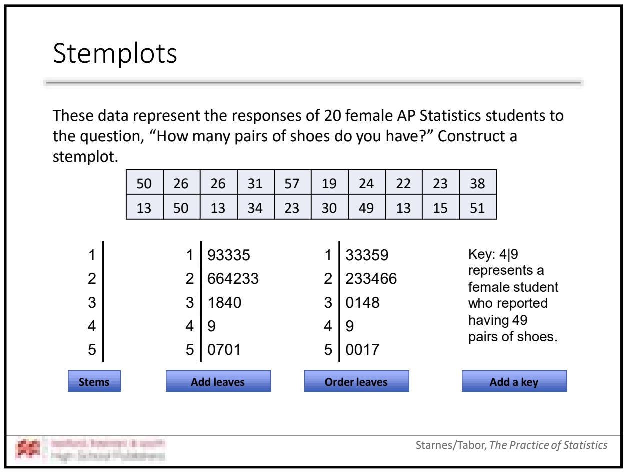
A **stemplot** shows each data value separated into two parts: a stem, which consists of all but the final digit, and a leaf, the final digit. The stems are ordered from lowest to highest and arranged in a vertical column. The leaves are arranged in increasing order out from the appropriate stems.

How to make a stemplot:

1) Separate each observation into a stem, consisting of all but the final digit, and a leaf, the final digit. Write the stems in a vertical column with the smallest at the top. Draw a vertical line at the right of this column.
2) Write each leaf in the row to the right of its stem.
3) Arrange the leaves in increasing order out from the stem.
4) Provide a key that identifies the variable and explains what the stems and leaves represent.

Starnes/Tabor, *The Practice of Statistics*

19



Stemplots

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

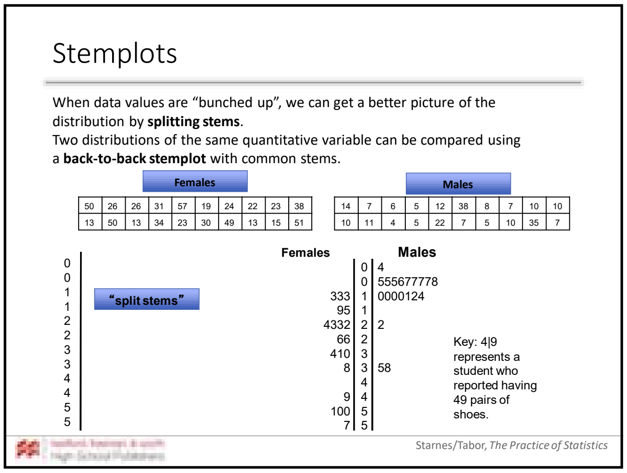
```
1 |        1 | 93335        1 | 33359
2 |        2 | 664233       2 | 233466
3 |        3 | 1840         3 | 0148
4 |        4 | 9            4 | 9
5 |        5 | 0701         5 | 0017
```

Key: 4|9 represents a female student who reported having 49 pairs of shoes.

Stems | Add leaves | Order leaves | Add a key

Starnes/Tabor, *The Practice of Statistics*

20



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

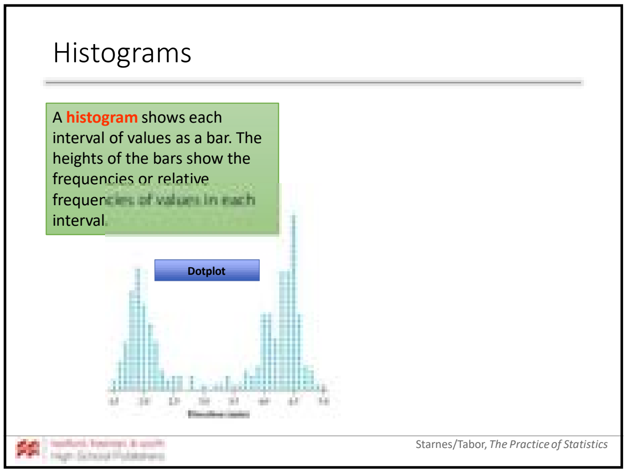
14	7	6	5	12	38	8	7	10	10
10	11	4	5	22	7	5	10	35	7

```
0 |                 Females      Males
0 |                         | 0 | 4
1 |                         | 0 | 555677778
1 |  “split stems”      333 | 1 | 0000124
2 |                      95 | 1 |
2 |                    4332 | 2 | 2
3 |                      66 | 2 |
3 |                     410 | 3 |
4 |                       8 | 3 | 58
4 |                         | 4 |
5 |                       9 | 4 |
5 |                     100 | 5 |
                          7 | 5 |
```

Key: 4|9 represents a student who reported having 49 pairs of shoes.

Starnes/Tabor, *The Practice of Statistics*

21

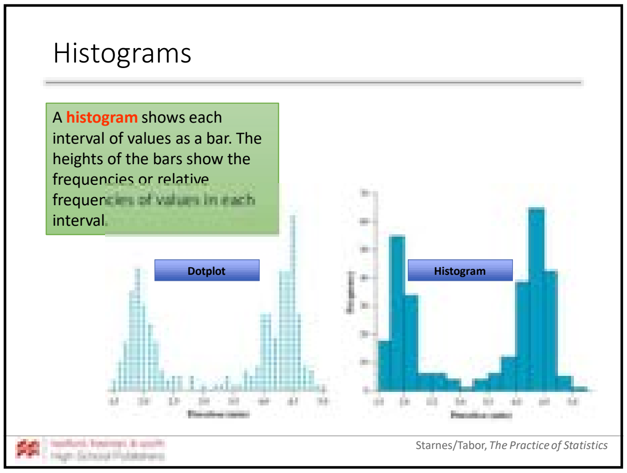


Histograms

A **histogram** shows each interval of values as a bar. The heights of the bars show the frequencies or relative frequencies of values in each interval.

Starnes/Tabor, *The Practice of Statistics*

22

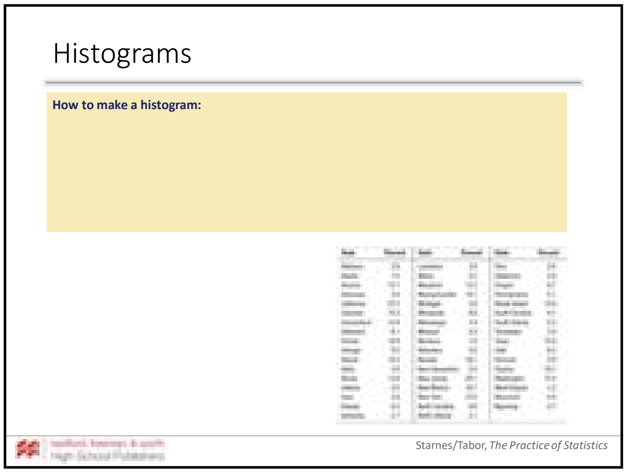


Histograms

A **histogram** shows each interval of values as a bar. The heights of the bars show the frequencies or relative frequencies of values in each interval.

Starnes/Tabor, *The Practice of Statistics*

23

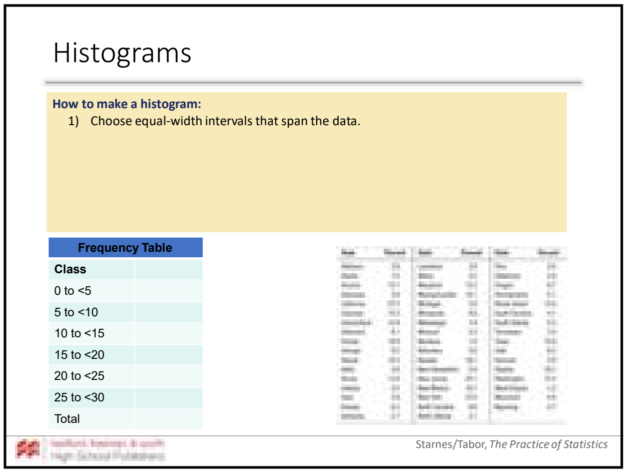


Histograms

How to make a histogram:

Starnes/Tabor, *The Practice of Statistics*

24



Histograms

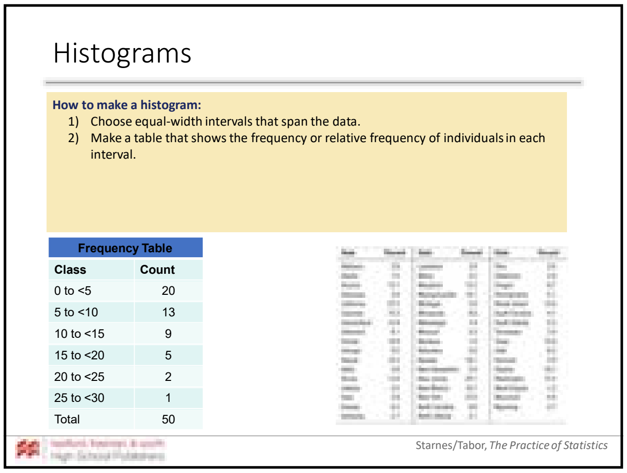
How to make a histogram:

1) Choose equal-width intervals that span the data.

Frequency Table	
Class	
0 to <5	
5 to <10	
10 to <15	
15 to <20	
20 to <25	
25 to <30	
Total	

Starnes/Tabor, *The Practice of Statistics*

25



Histograms

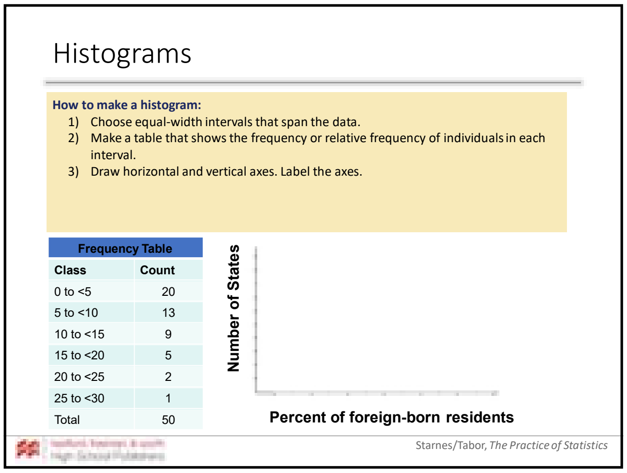
How to make a histogram:

1) Choose equal-width intervals that span the data.
2) Make a table that shows the frequency or relative frequency of individuals in each interval.

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

Starnes/Tabor, *The Practice of Statistics*

26



Histograms

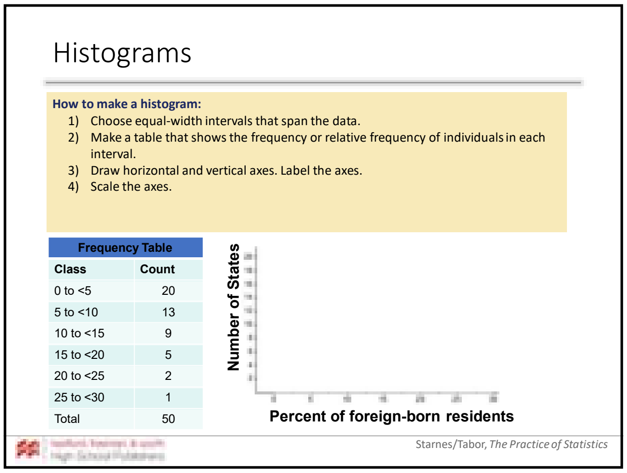
How to make a histogram:

1) Choose equal-width intervals that span the data.
2) Make a table that shows the frequency or relative frequency of individuals in each interval.
3) Draw horizontal and vertical axes. Label the axes.

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

Starnes/Tabor, *The Practice of Statistics*

27



Histograms

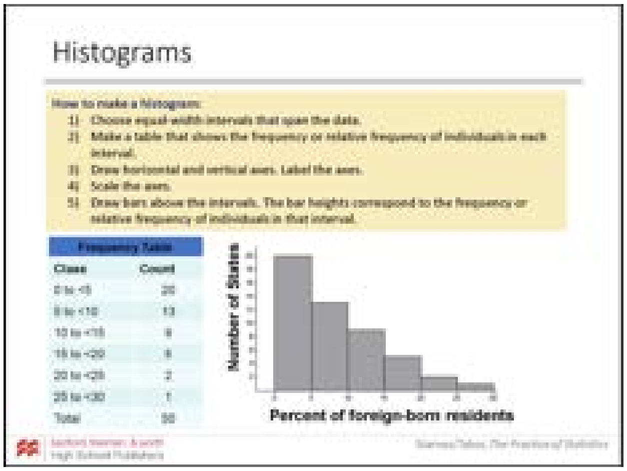
How to make a histogram:

1) Choose equal-width intervals that span the data.
2) Make a table that shows the frequency or relative frequency of individuals in each interval.
3) Draw horizontal and vertical axes. Label the axes.
4) Scale the axes.

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

Starnes/Tabor, *The Practice of Statistics*

28



Histograms

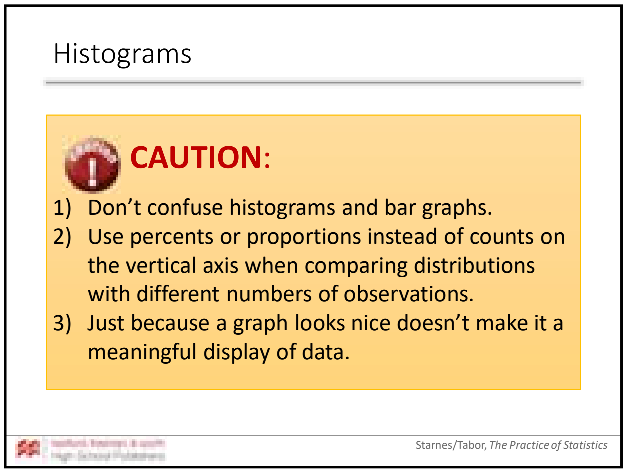
How to make a histogram:

1) Choose equal-width intervals that span the data.
2) Make a table that shows the frequency or relative frequency of individuals in each interval.
3) Draw horizontal and vertical axes. Label the axes.
4) Scale the axes.
5) Draw bars above the intervals. The bar heights correspond to the frequency or relative frequency of individuals in that interval.

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

Starnes/Tabor, *The Practice of Statistics*

29



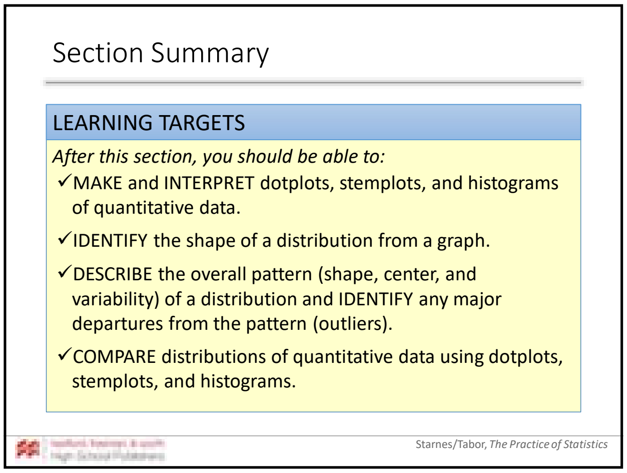
Histograms

CAUTION:

1) Don’t confuse histograms and bar graphs.
2) Use percents or proportions instead of counts on the vertical axis when comparing distributions with different numbers of observations.
3) Just because a graph looks nice doesn’t make it a meaningful display of data.

Starnes/Tabor, *The Practice of Statistics*

30



Section Summary

LEARNING TARGETS

After this section, you should be able to:

✓MAKE and INTERPRET dotplots, stemplots, and histograms of quantitative data.

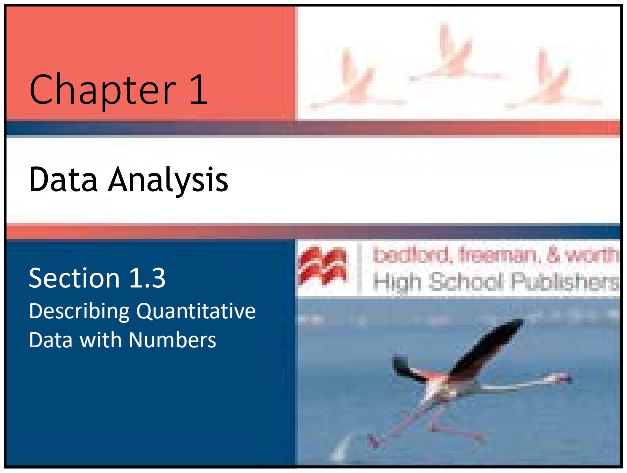
✓IDENTIFY the shape of a distribution from a graph.

✓DESCRIBE the overall pattern (shape, center, and variability) of a distribution and IDENTIFY any major departures from the pattern (outliers).

✓COMPARE distributions of quantitative data using dotplots, stemplots, and histograms.

Starnes/Tabor, *The Practice of Statistics*

31

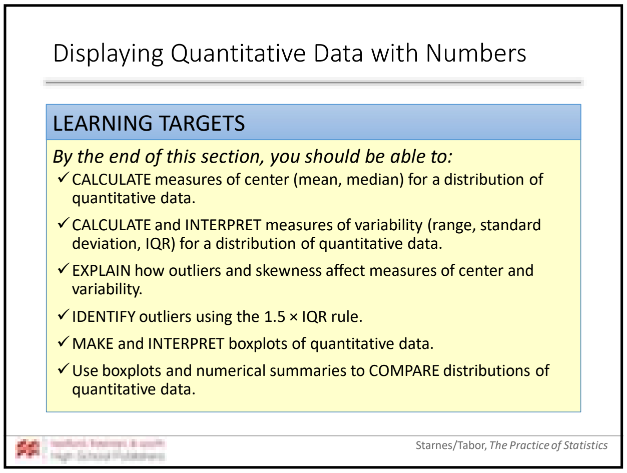


Chapter 1

Data Analysis

Section 1.3
Describing Quantitative Data with Numbers

1



Displaying Quantitative Data with Numbers

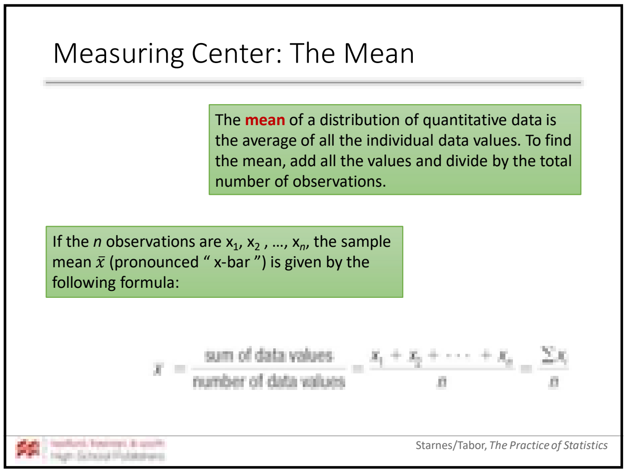
LEARNING TARGETS

By the end of this section, you should be able to:

- ✓ CALCULATE measures of center (mean, median) for a distribution of quantitative data.
- ✓ CALCULATE and INTERPRET measures of variability (range, standard deviation, IQR) for a distribution of quantitative data.
- ✓ EXPLAIN how outliers and skewness affect measures of center and variability.
- ✓ IDENTIFY outliers using the 1.5 × IQR rule.
- ✓ MAKE and INTERPRET boxplots of quantitative data.
- ✓ Use boxplots and numerical summaries to COMPARE distributions of quantitative data.

Starnes/Tabor, *The Practice of Statistics*

2



Measuring Center: The Mean

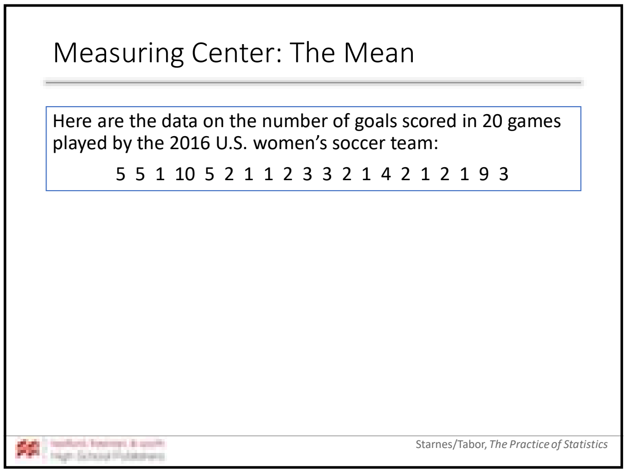
The **mean** of a distribution of quantitative data is the average of all the individual data values. To find the mean, add all the values and divide by the total number of observations.

If the n observations are x_1, x_2 , …, x_n, the sample mean $\bar{x}$ (pronounced " x-bar ") is given by the following formula:

$$\bar{x} = \frac{\text{sum of data values}}{\text{number of data values}} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x_i}{n}$$

Starnes/Tabor, *The Practice of Statistics*

3



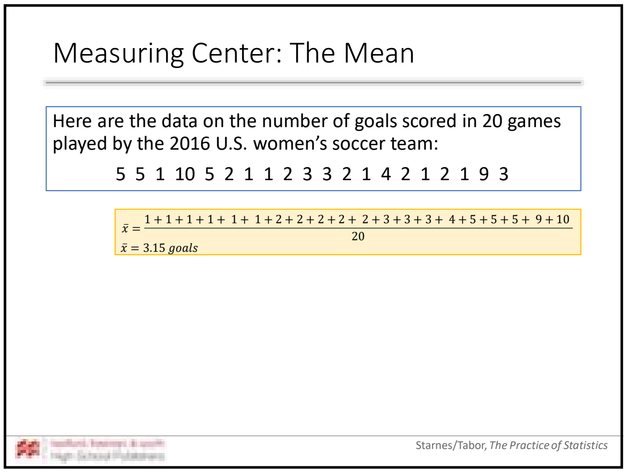
Measuring Center: The Mean

Here are the data on the number of goals scored in 20 games played by the 2016 U.S. women's soccer team:

5 5 1 10 5 2 1 1 2 3 3 2 1 4 2 1 2 1 9 3

Starnes/Tabor, *The Practice of Statistics*

4



Measuring Center: The Mean

Here are the data on the number of goals scored in 20 games played by the 2016 U.S. women's soccer team:

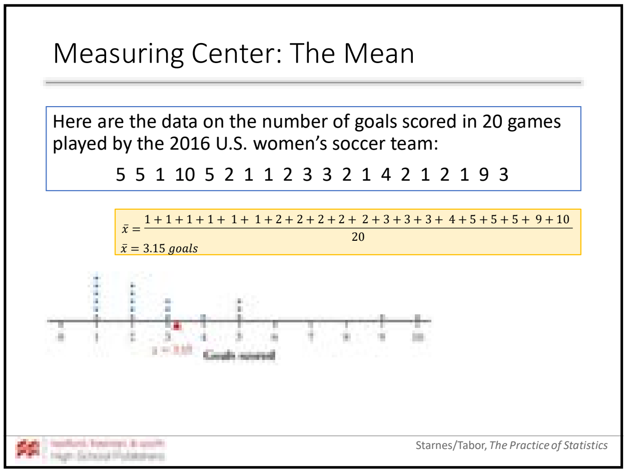
5 5 1 10 5 2 1 1 2 3 3 2 1 4 2 1 2 1 9 3

$$\bar{x} = \frac{1+1+1+1+1+1+2+2+2+2+2+3+3+3+4+5+5+5+9+10}{20}$$

$$\bar{x} = 3.15\ goals$$

Starnes/Tabor, *The Practice of Statistics*

5



Measuring Center: The Mean

Here are the data on the number of goals scored in 20 games played by the 2016 U.S. women's soccer team:

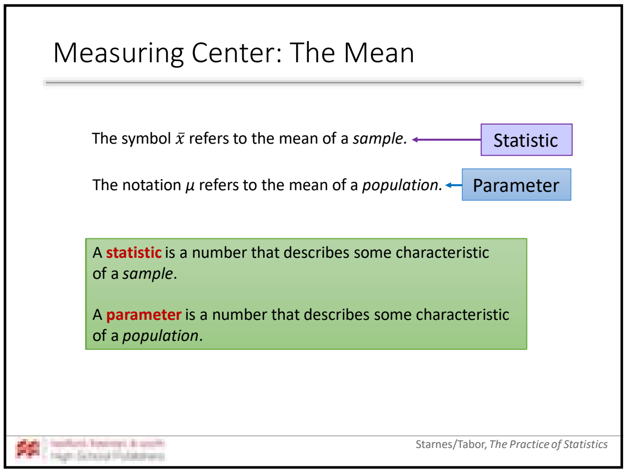
5 5 1 10 5 2 1 1 2 3 3 2 1 4 2 1 2 1 9 3

$$\bar{x} = \frac{1+1+1+1+1+1+2+2+2+2+2+3+3+3+4+5+5+5+9+10}{20}$$

$$\bar{x} = 3.15\ goals$$

Starnes/Tabor, *The Practice of Statistics*

6



Measuring Center: The Mean

The symbol $\bar{x}$ refers to the mean of a *sample*. ← Statistic

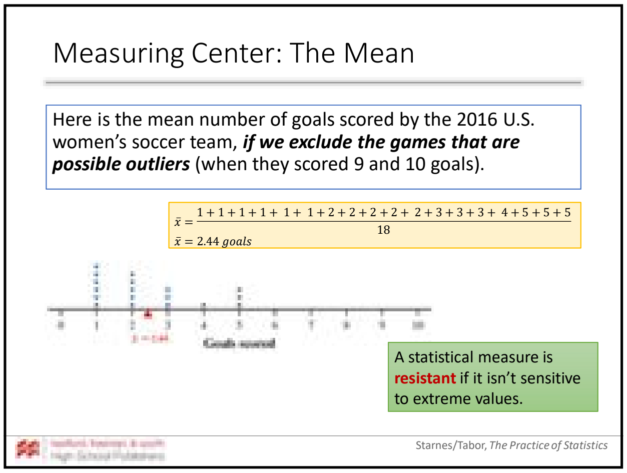
The notation μ refers to the mean of a *population*. ← Parameter

A **statistic** is a number that describes some characteristic of a *sample*.

A **parameter** is a number that describes some characteristic of a *population*.

Starnes/Tabor, *The Practice of Statistics*

7



Measuring Center: The Mean

Here is the mean number of goals scored by the 2016 U.S. women's soccer team, ***if we exclude the games that are possible outliers*** (when they scored 9 and 10 goals).

$$\bar{x} = \frac{1+1+1+1+\ 1+\ 1+2+2+2+2+\ 2+3+3+3+\ 4+5+5+5}{18}$$

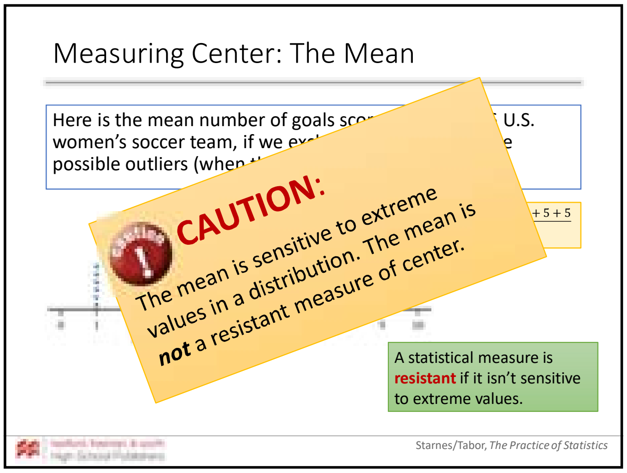
$$\bar{x} = 2.44\ goals$$

0
1
2
3
4
5
6
7
8
9
10
$\bar{x} = 2.44$
Goals scored

A statistical measure is **resistant** if it isn't sensitive to extreme values.

Starnes/Tabor, *The Practice of Statistics*

8



Measuring Center: The Mean

Here is the mean number of goals sco… U.S. women's soccer team, if we ex… e possible outliers (when t…

+ 5 + 5

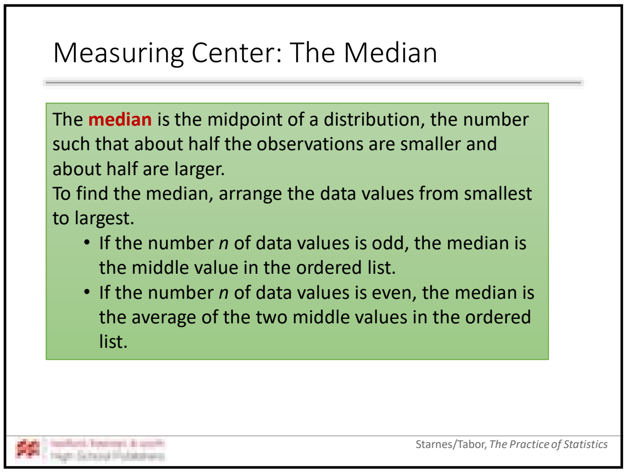
CAUTION:

The mean is sensitive to extreme values in a distribution. The mean is ***not*** a resistant measure of center.

A statistical measure is **resistant** if it isn't sensitive to extreme values.

Starnes/Tabor, *The Practice of Statistics*

9



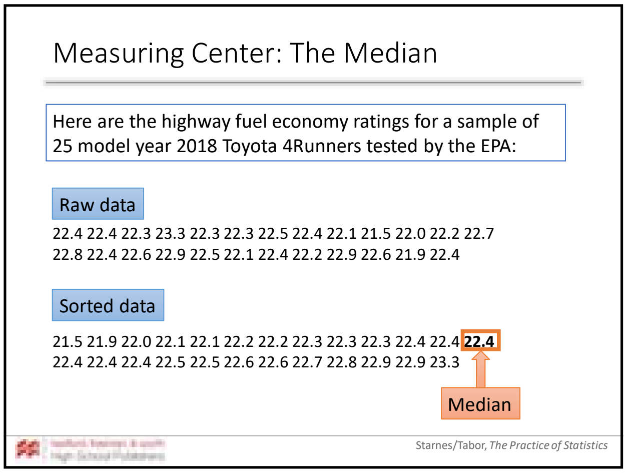
Measuring Center: The Median

The **median** is the midpoint of a distribution, the number such that about half the observations are smaller and about half are larger.
To find the median, arrange the data values from smallest to largest.

- If the number *n* of data values is odd, the median is the middle value in the ordered list.
- If the number *n* of data values is even, the median is the average of the two middle values in the ordered list.

Starnes/Tabor, *The Practice of Statistics*

10



Measuring Center: The Median

Here are the highway fuel economy ratings for a sample of 25 model year 2018 Toyota 4Runners tested by the EPA:

Raw data

22.4 22.4 22.3 23.3 22.3 22.3 22.5 22.4 22.1 21.5 22.0 22.2 22.7
22.8 22.4 22.6 22.9 22.5 22.1 22.4 22.2 22.9 22.6 21.9 22.4

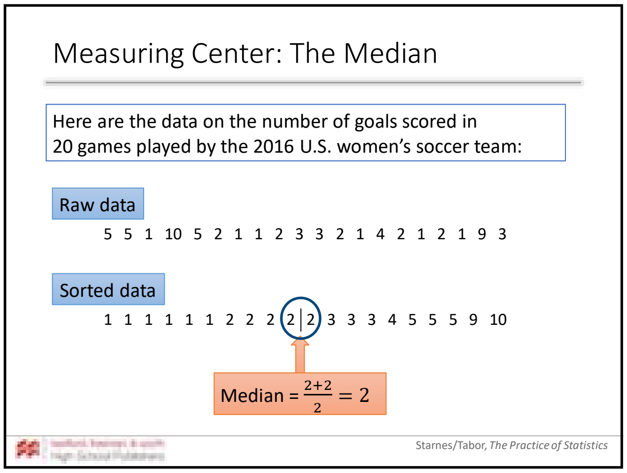
Sorted data

21.5 21.9 22.0 22.1 22.1 22.2 22.2 22.3 22.3 22.3 22.4 22.4 **22.4**
22.4 22.4 22.4 22.5 22.5 22.6 22.6 22.7 22.8 22.9 22.9 23.3

Median

Starnes/Tabor, *The Practice of Statistics*

11



Measuring Center: The Median

Here are the data on the number of goals scored in 20 games played by the 2016 U.S. women's soccer team:

Raw data

5 5 1 10 5 2 1 1 2 3 3 2 1 4 2 1 2 1 9 3

Sorted data

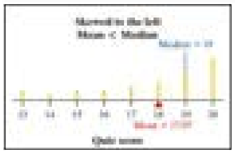
1 1 1 1 1 1 2 2 2 2 | 2 3 3 3 4 5 5 5 9 10

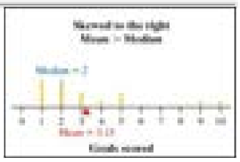
Median = $\frac{2+2}{2} = 2$

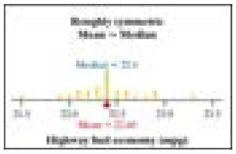
Starnes/Tabor, *The Practice of Statistics*

12

Comparing the Mean and the Median







Starnes/Tabor, *The Practice of Statistics*

13

Comparing the Mean and the Median

Effect of Skewness and Outliers on Measures of Center

- If a distribution of quantitative data is roughly symmetric and has no outliers, the mean and median will be similar.
- If the distribution is strongly skewed, the mean will be pulled in the direction of the skewness but the median won't. For a right-skewed distribution, we expect the mean to be greater than the median. For a left-skewed distribution, we expect the mean to be less than the median.
- The median is resistant to outliers but the mean isn't.

Starnes/Tabor, *The Practice of Statistics*

14

Measuring Variability: The Range

The **range** of a distribution is the distance between the minimum value and the maximum value. That is,

Range = Maximum – Minimum

Here are the data on the number of goals scored in 20 games played by the 2016 U.S. women's soccer team:

5 5 1 10 5 2 1 1 2 3 3 2 1 4 2 1 2 1 9 3

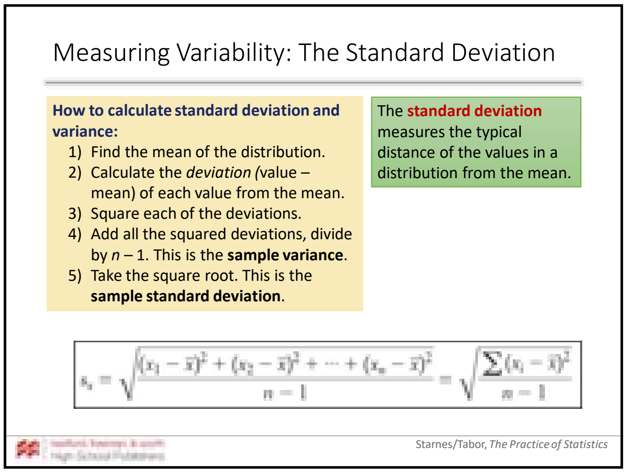
$\text{Range} = 10 \ - 1 = 9 \text{ goals}$

CAUTION:

- The range of a data set is a single number.
- The range is ***not*** a resistant measure of variability.

Starnes/Tabor, *The Practice of Statistics*

15



Measuring Variability: The Standard Deviation

How to calculate standard deviation and variance:

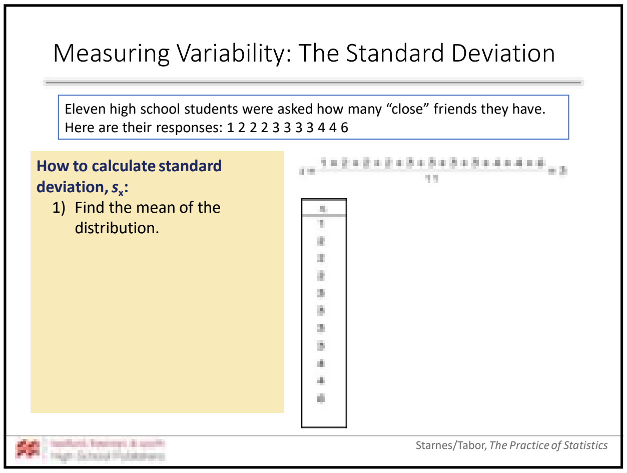
1) Find the mean of the distribution.
2) Calculate the *deviation* (value – mean) of each value from the mean.
3) Square each of the deviations.
4) Add all the squared deviations, divide by *n* – 1. This is the **sample variance**.
5) Take the square root. This is the **sample standard deviation**.

The **standard deviation** measures the typical distance of the values in a distribution from the mean.

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

Starnes/Tabor, *The Practice of Statistics*

16



Measuring Variability: The Standard Deviation

Eleven high school students were asked how many “close” friends they have. Here are their responses: 1 2 2 2 3 3 3 3 4 4 6

How to calculate standard deviation, s_x:

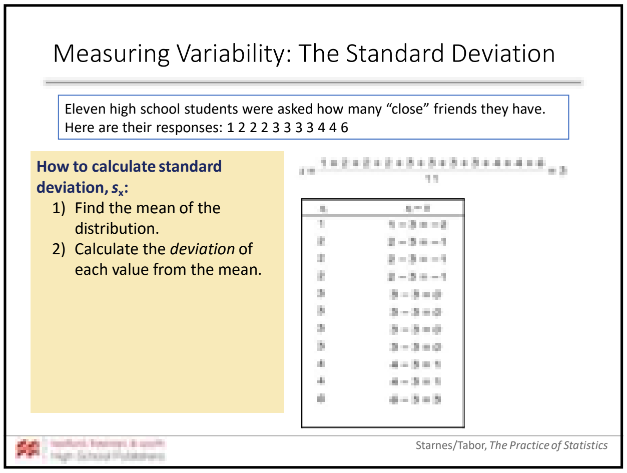
1) Find the mean of the distribution.

$$\bar{x} = \frac{1+2+2+2+3+3+3+3+4+4+6}{11} = 3$$

x_i
1
2
2
2
3
3
3
3
4
4
6

Starnes/Tabor, *The Practice of Statistics*

17



Measuring Variability: The Standard Deviation

Eleven high school students were asked how many “close” friends they have. Here are their responses: 1 2 2 2 3 3 3 3 4 4 6

How to calculate standard deviation, s_x:

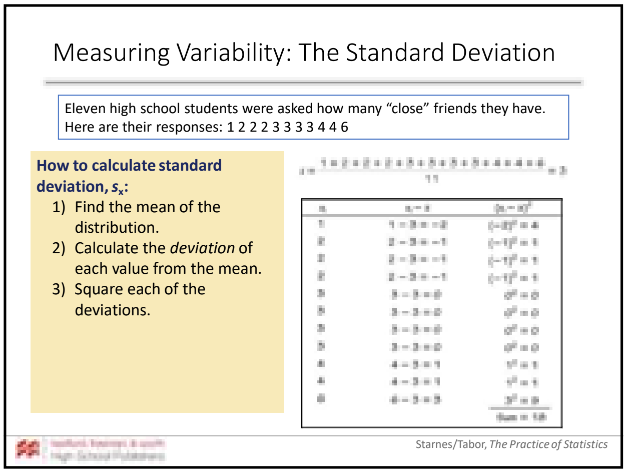
1) Find the mean of the distribution.
2) Calculate the *deviation* of each value from the mean.

$$\bar{x} = \frac{1+2+2+2+3+3+3+3+4+4+6}{11} = 3$$

x_i	$x_i - \bar{x}$
1	$1-3=-2$
2	$2-3=-1$
2	$2-3=-1$
2	$2-3=-1$
3	$3-3=0$
3	$3-3=0$
3	$3-3=0$
3	$3-3=0$
4	$4-3=1$
4	$4-3=1$
6	$6-3=3$

Starnes/Tabor, *The Practice of Statistics*

18



Measuring Variability: The Standard Deviation

Eleven high school students were asked how many "close" friends they have. Here are their responses: 1 2 2 2 3 3 3 3 4 4 6

How to calculate standard deviation, s_x:

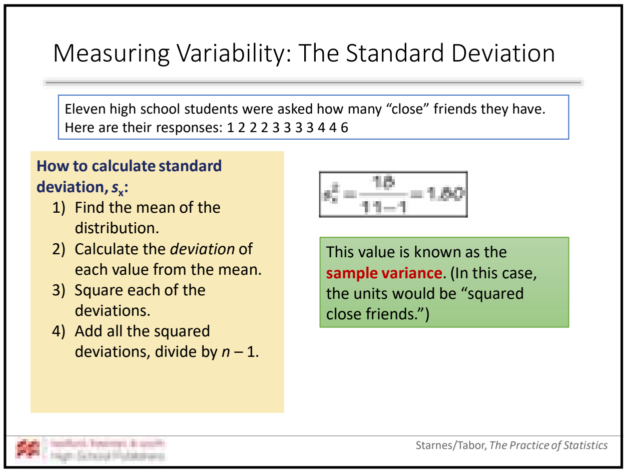
1) Find the mean of the distribution.
2) Calculate the *deviation* of each value from the mean.
3) Square each of the deviations.

$$\bar{x} = \frac{1+2+2+2+3+3+3+3+4+4+6}{11} = 3$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	$1 - 3 = -2$	$(-2)^2 = 4$
2	$2 - 3 = -1$	$(-1)^2 = 1$
2	$2 - 3 = -1$	$(-1)^2 = 1$
2	$2 - 3 = -1$	$(-1)^2 = 1$
3	$3 - 3 = 0$	$0^2 = 0$
3	$3 - 3 = 0$	$0^2 = 0$
3	$3 - 3 = 0$	$0^2 = 0$
3	$3 - 3 = 0$	$0^2 = 0$
4	$4 - 3 = 1$	$1^2 = 1$
4	$4 - 3 = 1$	$1^2 = 1$
6	$6 - 3 = 3$	$3^2 = 9$
		Sum = 18

Starnes/Tabor, *The Practice of Statistics*

19



Measuring Variability: The Standard Deviation

Eleven high school students were asked how many "close" friends they have. Here are their responses: 1 2 2 2 3 3 3 3 4 4 6

How to calculate standard deviation, s_x:

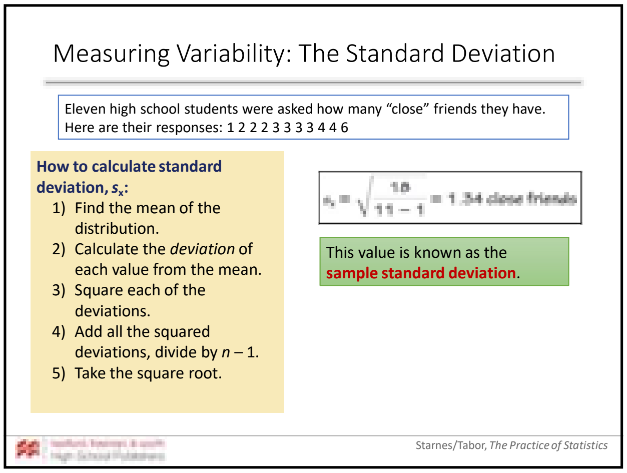
1) Find the mean of the distribution.
2) Calculate the *deviation* of each value from the mean.
3) Square each of the deviations.
4) Add all the squared deviations, divide by *n* – 1.

$$s_x^2 = \frac{18}{11-1} = 1.80$$

This value is known as the **sample variance**. (In this case, the units would be "squared close friends.")

Starnes/Tabor, *The Practice of Statistics*

20



Measuring Variability: The Standard Deviation

Eleven high school students were asked how many "close" friends they have. Here are their responses: 1 2 2 2 3 3 3 3 4 4 6

How to calculate standard deviation, s_x:

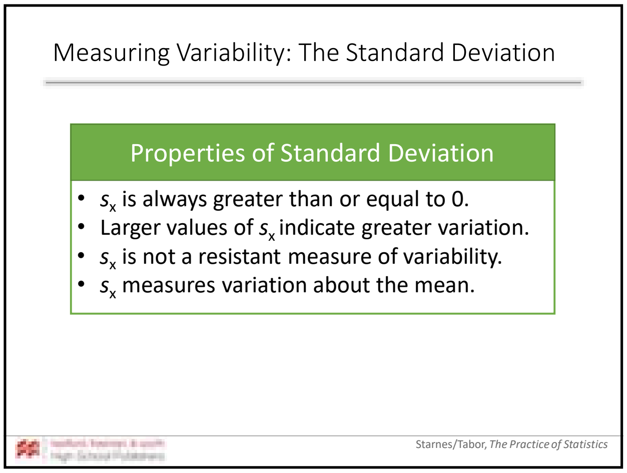
1) Find the mean of the distribution.
2) Calculate the *deviation* of each value from the mean.
3) Square each of the deviations.
4) Add all the squared deviations, divide by *n* – 1.
5) Take the square root.

$$s_x = \sqrt{\frac{18}{11-1}} = 1.34 \text{ close friends}$$

This value is known as the **sample standard deviation**.

Starnes/Tabor, *The Practice of Statistics*

21



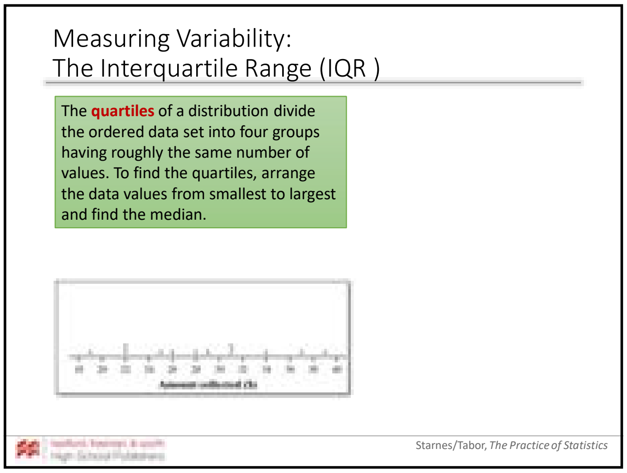
Measuring Variability: The Standard Deviation

Properties of Standard Deviation

- s_x is always greater than or equal to 0.
- Larger values of s_x indicate greater variation.
- s_x is not a resistant measure of variability.
- s_x measures variation about the mean.

Starnes/Tabor, *The Practice of Statistics*

22



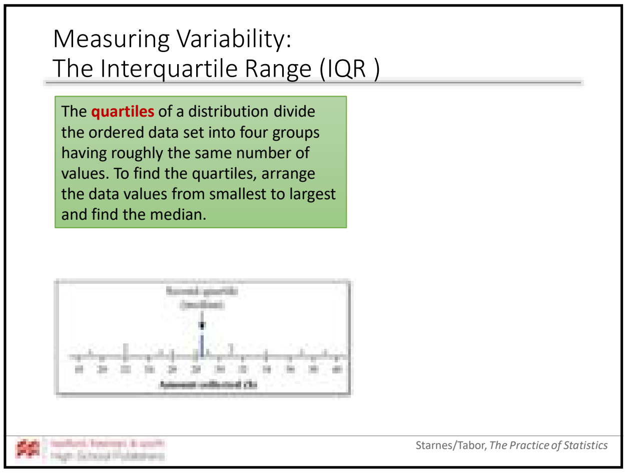
Measuring Variability: The Interquartile Range (IQR)

The **quartiles** of a distribution divide the ordered data set into four groups having roughly the same number of values. To find the quartiles, arrange the data values from smallest to largest and find the median.

Amount collected ($)

Starnes/Tabor, *The Practice of Statistics*

23



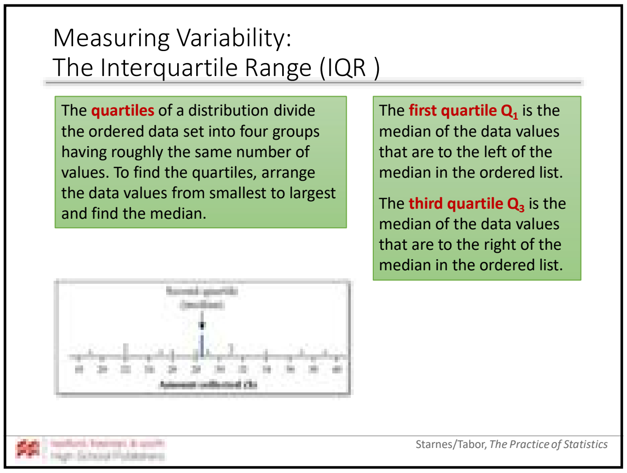
Measuring Variability: The Interquartile Range (IQR)

The **quartiles** of a distribution divide the ordered data set into four groups having roughly the same number of values. To find the quartiles, arrange the data values from smallest to largest and find the median.

Second quartile
(median)
Amount collected ($)

Starnes/Tabor, *The Practice of Statistics*

24



Measuring Variability: The Interquartile Range (IQR)

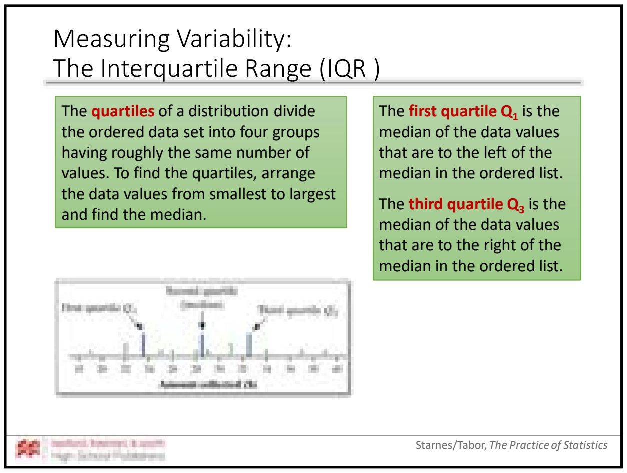
The **quartiles** of a distribution divide the ordered data set into four groups having roughly the same number of values. To find the quartiles, arrange the data values from smallest to largest and find the median.

The **first quartile Q_1** is the median of the data values that are to the left of the median in the ordered list.

The **third quartile Q_3** is the median of the data values that are to the right of the median in the ordered list.

Starnes/Tabor, *The Practice of Statistics*

25



Measuring Variability: The Interquartile Range (IQR)

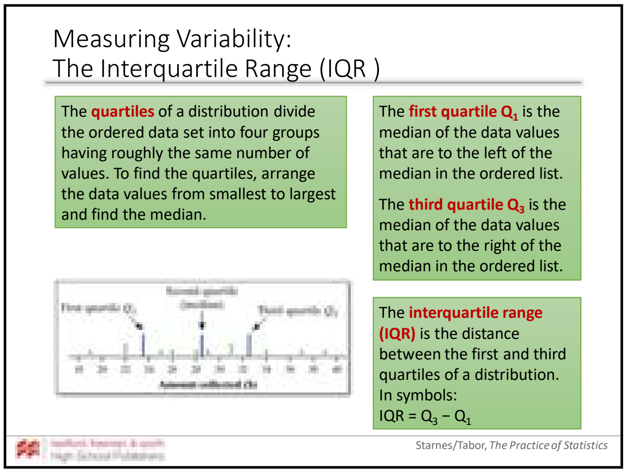
The **quartiles** of a distribution divide the ordered data set into four groups having roughly the same number of values. To find the quartiles, arrange the data values from smallest to largest and find the median.

The **first quartile Q_1** is the median of the data values that are to the left of the median in the ordered list.

The **third quartile Q_3** is the median of the data values that are to the right of the median in the ordered list.

Starnes/Tabor, *The Practice of Statistics*

26



Measuring Variability: The Interquartile Range (IQR)

The **quartiles** of a distribution divide the ordered data set into four groups having roughly the same number of values. To find the quartiles, arrange the data values from smallest to largest and find the median.

The **first quartile Q_1** is the median of the data values that are to the left of the median in the ordered list.

The **third quartile Q_3** is the median of the data values that are to the right of the median in the ordered list.

The **interquartile range (IQR)** is the distance between the first and third quartiles of a distribution. In symbols: IQR = $Q_3 - Q_1$

Starnes/Tabor, *The Practice of Statistics*

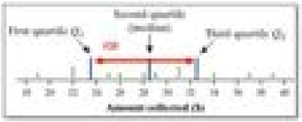
27

Measuring Variability: The Interquartile Range (IQR)

The **quartiles** of a distribution divide the ordered data set into four groups having roughly the same number of values. To find the quartiles, arrange the data values from smallest to largest and find the median.

The **first quartile Q_1** is the median of the data values that are to the left of the median in the ordered list.

The **third quartile Q_3** is the median of the data values that are to the right of the median in the ordered list.



The **interquartile range (IQR)** is the distance between the first and third quartiles of a distribution. In symbols:
IQR = $Q_3 - Q_1$

Starnes/Tabor, *The Practice of Statistics*

28

Measuring Variability: The Interquartile Range (IQR)

Travel times for 20 New Yorkers:

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

Starnes/Tabor, *The Practice of Statistics*

29

Measuring Variability: The Interquartile Range (IQR)

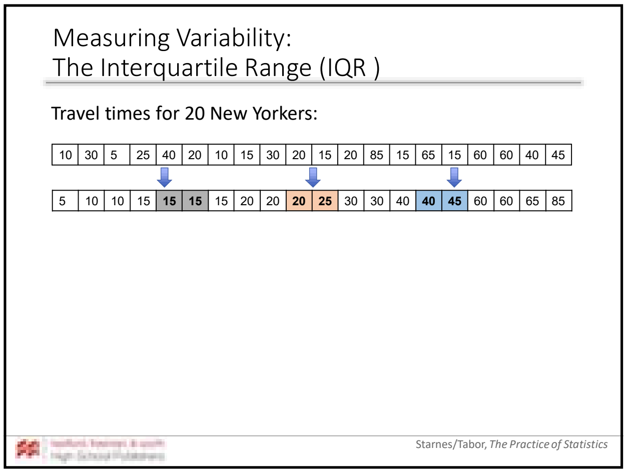
Travel times for 20 New Yorkers:

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

Starnes/Tabor, *The Practice of Statistics*

30



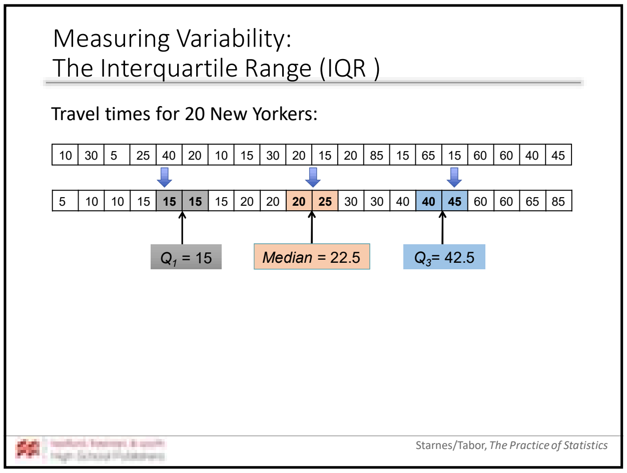
Measuring Variability: The Interquartile Range (IQR)

Travel times for 20 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

Starnes/Tabor, *The Practice of Statistics*

31



Measuring Variability: The Interquartile Range (IQR)

Travel times for 20 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

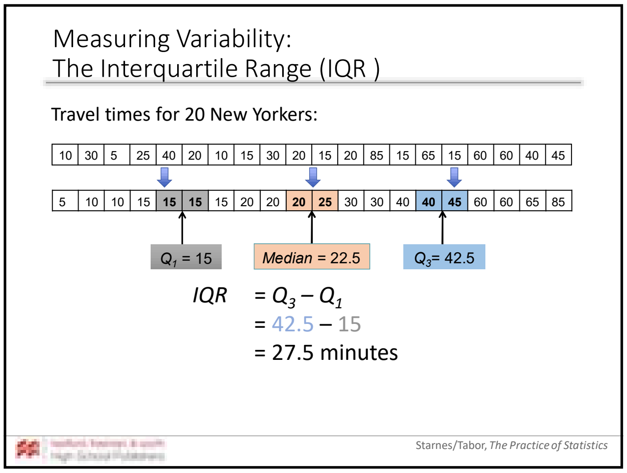
Q_1 = 15

Median = 22.5

Q_3= 42.5

Starnes/Tabor, *The Practice of Statistics*

32



Measuring Variability: The Interquartile Range (IQR)

Travel times for 20 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

Median = 22.5

Q_3= 42.5

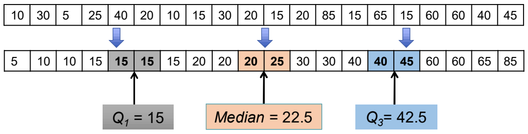
$$\begin{aligned} IQR &= Q_3 - Q_1 \\ &= 42.5 - 15 \\ &= 27.5 \text{ minutes} \end{aligned}$$

Starnes/Tabor, *The Practice of Statistics*

33

Measuring Variability: The Interquartile Range (IQR)

Travel times for 20 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

Median = 22.5

Q_3= 42.5

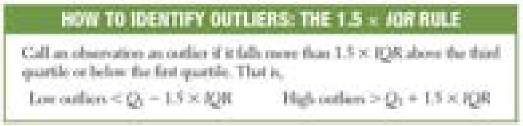
$$\begin{aligned} IQR &= Q_3 - Q_1 \\ &= 42.5 - 15 \\ &= 27.5 \text{ minutes} \end{aligned}$$

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

34

Identifying Outliers

Although there are several rules for outliers, one of the most common rules is the 1.5 × IQR rule.



HOW TO IDENTIFY OUTLIERS: THE 1.5 × IQR RULE

Call an observation an outlier if it falls more than 1.5 × IQR above the third quartile or below the first quartile. That is,

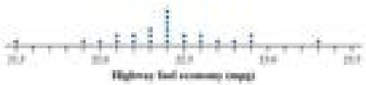
Low outliers < Q_1 − 1.5 × IQR   High outliers > Q_3 + 1.5 × IQR

35

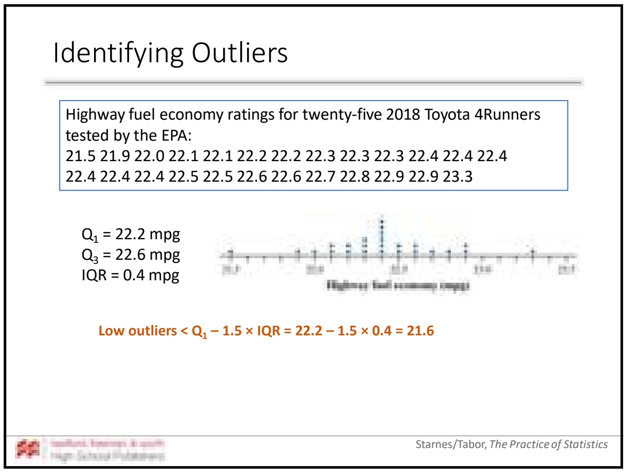
Identifying Outliers

Highway fuel economy ratings for twenty-five 2018 Toyota 4Runners tested by the EPA:
21.5 21.9 22.0 22.1 22.1 22.2 22.2 22.3 22.3 22.3 22.4 22.4 22.4
22.4 22.4 22.4 22.5 22.5 22.6 22.6 22.7 22.8 22.9 22.9 23.3

Q_1 = 22.2 mpg
Q_3 = 22.6 mpg
IQR = 0.4 mpg



36



Identifying Outliers

Highway fuel economy ratings for twenty-five 2018 Toyota 4Runners tested by the EPA:
21.5 21.9 22.0 22.1 22.1 22.2 22.2 22.3 22.3 22.3 22.4 22.4 22.4
22.4 22.4 22.4 22.5 22.5 22.6 22.6 22.7 22.8 22.9 22.9 23.3

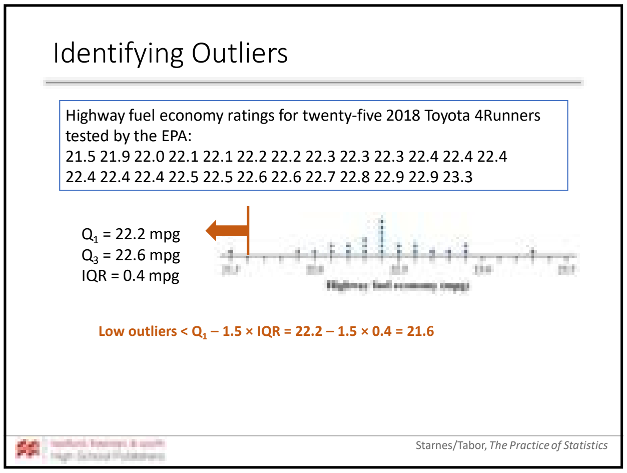
Q_1 = 22.2 mpg
Q_3 = 22.6 mpg
IQR = 0.4 mpg

21.5
22.0
22.5
23.0
23.5
Highway fuel economy (mpg)

Low outliers < Q_1 – 1.5 × IQR = 22.2 – 1.5 × 0.4 = 21.6

Starnes/Tabor, *The Practice of Statistics*

37



Identifying Outliers

Highway fuel economy ratings for twenty-five 2018 Toyota 4Runners tested by the EPA:
21.5 21.9 22.0 22.1 22.1 22.2 22.2 22.3 22.3 22.3 22.4 22.4 22.4
22.4 22.4 22.4 22.5 22.5 22.6 22.6 22.7 22.8 22.9 22.9 23.3

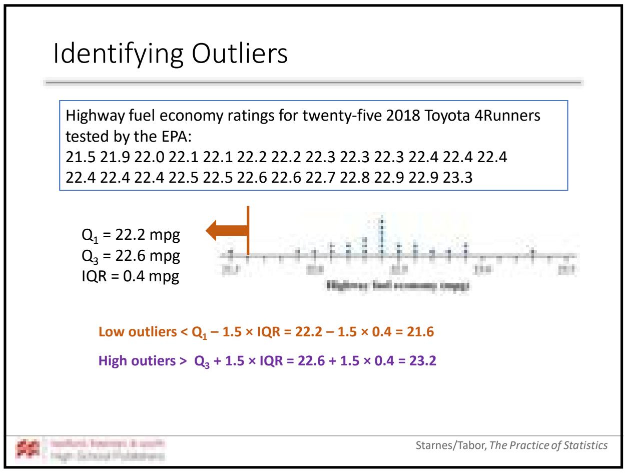
Q_1 = 22.2 mpg
Q_3 = 22.6 mpg
IQR = 0.4 mpg

21.5
22.0
22.5
23.0
23.5
Highway fuel economy (mpg)

Low outliers < Q_1 – 1.5 × IQR = 22.2 – 1.5 × 0.4 = 21.6

Starnes/Tabor, *The Practice of Statistics*

38



Identifying Outliers

Highway fuel economy ratings for twenty-five 2018 Toyota 4Runners tested by the EPA:
21.5 21.9 22.0 22.1 22.1 22.2 22.2 22.3 22.3 22.3 22.4 22.4 22.4
22.4 22.4 22.4 22.5 22.5 22.6 22.6 22.7 22.8 22.9 22.9 23.3

Q_1 = 22.2 mpg
Q_3 = 22.6 mpg
IQR = 0.4 mpg

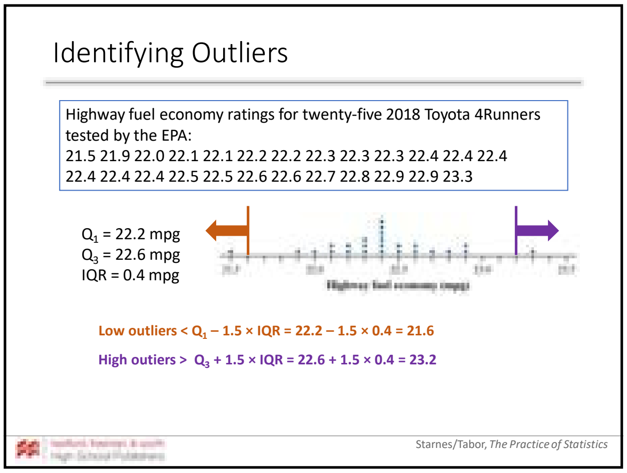
21.5
22.0
22.5
23.0
23.5
Highway fuel economy (mpg)

Low outliers < Q_1 – 1.5 × IQR = 22.2 – 1.5 × 0.4 = 21.6

High outiers > Q_3 + 1.5 × IQR = 22.6 + 1.5 × 0.4 = 23.2

Starnes/Tabor, *The Practice of Statistics*

39



Identifying Outliers

Highway fuel economy ratings for twenty-five 2018 Toyota 4Runners tested by the EPA:
21.5 21.9 22.0 22.1 22.1 22.2 22.2 22.3 22.3 22.3 22.4 22.4 22.4
22.4 22.4 22.4 22.5 22.5 22.6 22.6 22.7 22.8 22.9 22.9 23.3

Q_1 = 22.2 mpg
Q_3 = 22.6 mpg
IQR = 0.4 mpg

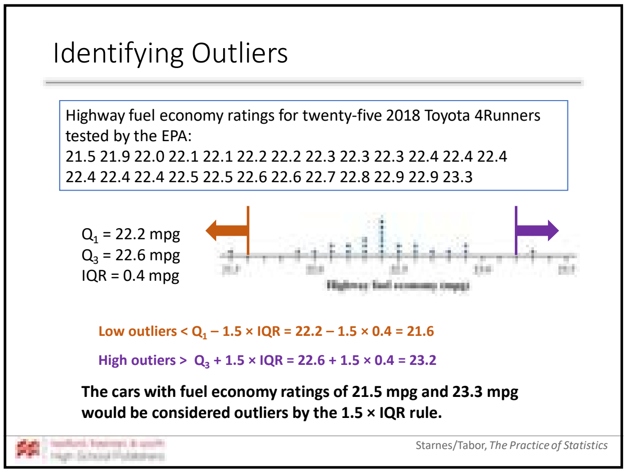
21.5
22.0
22.5
23.0
23.5
Highway fuel economy (mpg)

Low outliers < Q_1 – 1.5 × IQR = 22.2 – 1.5 × 0.4 = 21.6

High outiers > Q_3 + 1.5 × IQR = 22.6 + 1.5 × 0.4 = 23.2

Starnes/Tabor, *The Practice of Statistics*

40



Identifying Outliers

Highway fuel economy ratings for twenty-five 2018 Toyota 4Runners tested by the EPA:
21.5 21.9 22.0 22.1 22.1 22.2 22.2 22.3 22.3 22.3 22.4 22.4 22.4
22.4 22.4 22.4 22.5 22.5 22.6 22.6 22.7 22.8 22.9 22.9 23.3

Q_1 = 22.2 mpg
Q_3 = 22.6 mpg
IQR = 0.4 mpg

21.5
22.0
22.5
23.0
23.5
Highway fuel economy (mpg)

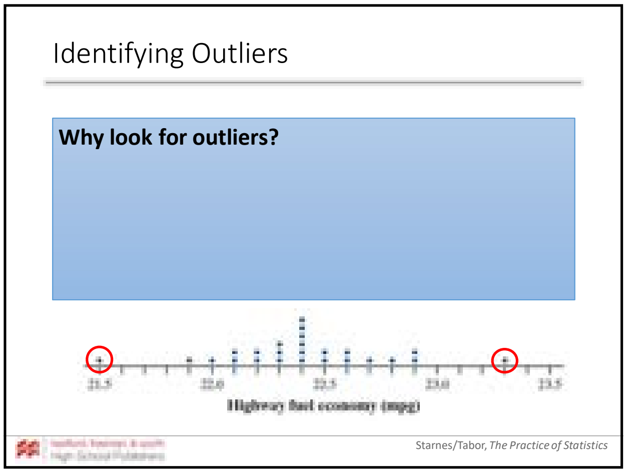
Low outliers < Q_1 – 1.5 × IQR = 22.2 – 1.5 × 0.4 = 21.6

High outiers > Q_3 + 1.5 × IQR = 22.6 + 1.5 × 0.4 = 23.2

The cars with fuel economy ratings of 21.5 mpg and 23.3 mpg would be considered outliers by the 1.5 × IQR rule.

Starnes/Tabor, *The Practice of Statistics*

41



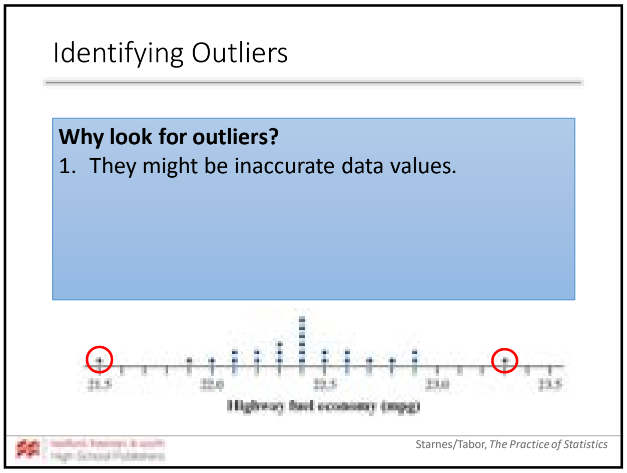
Identifying Outliers

Why look for outliers?

21.5
22.0
22.5
23.0
23.5
Highway fuel economy (mpg)

Starnes/Tabor, *The Practice of Statistics*

42



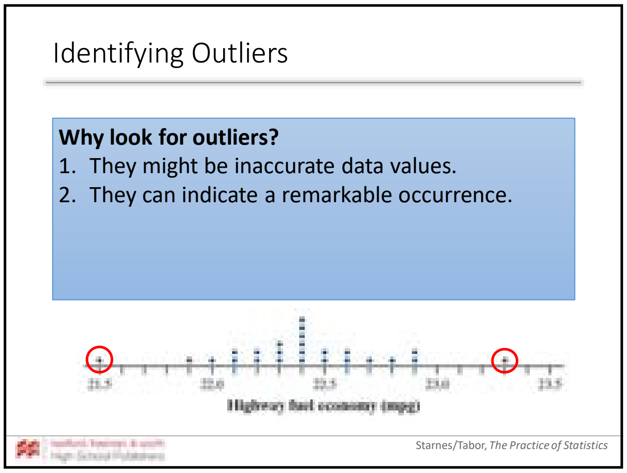
Identifying Outliers

Why look for outliers?

1. They might be inaccurate data values.

Starnes/Tabor, *The Practice of Statistics*

43



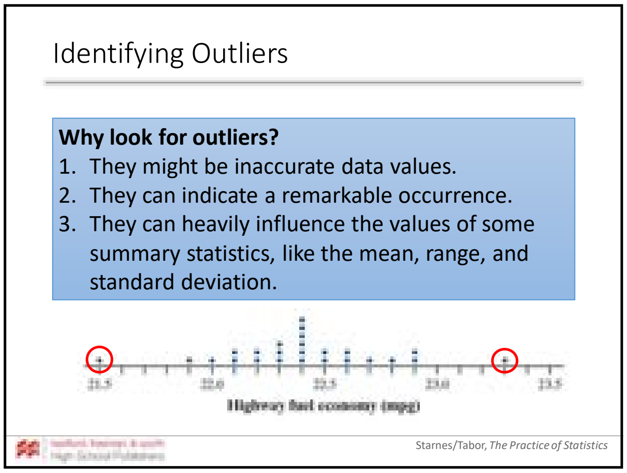
Identifying Outliers

Why look for outliers?

1. They might be inaccurate data values.
2. They can indicate a remarkable occurrence.

Starnes/Tabor, *The Practice of Statistics*

44



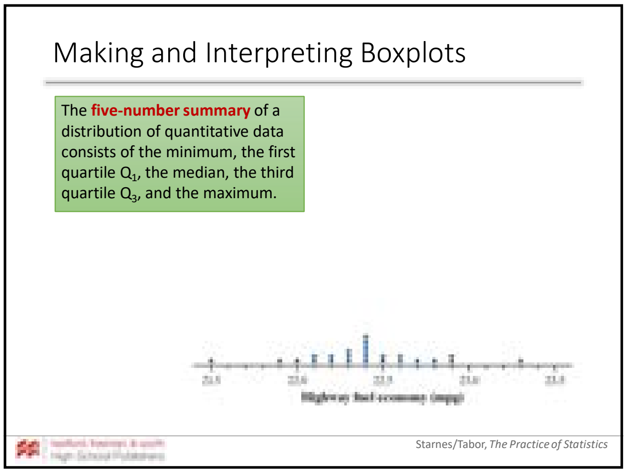
Identifying Outliers

Why look for outliers?

1. They might be inaccurate data values.
2. They can indicate a remarkable occurrence.
3. They can heavily influence the values of some summary statistics, like the mean, range, and standard deviation.

Starnes/Tabor, *The Practice of Statistics*

45

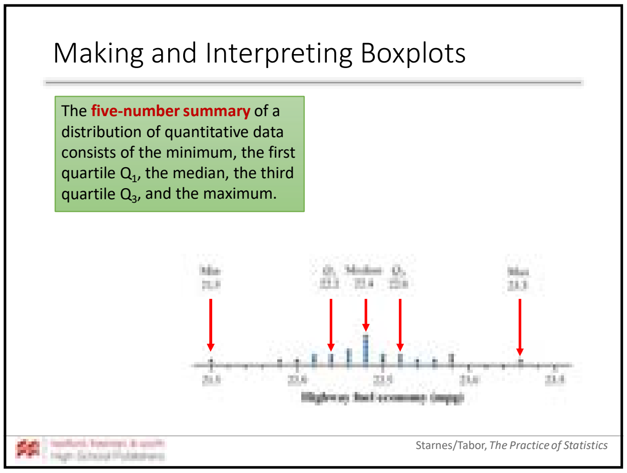


Making and Interpreting Boxplots

The **five-number summary** of a distribution of quantitative data consists of the minimum, the first quartile Q_1, the median, the third quartile Q_3, and the maximum.

Starnes/Tabor, *The Practice of Statistics*

46

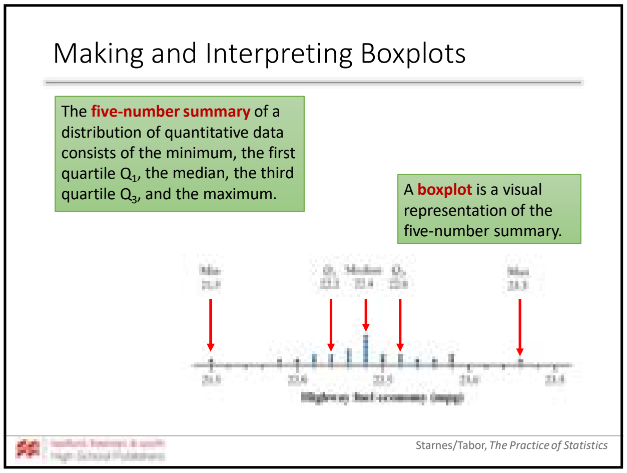


Making and Interpreting Boxplots

The **five-number summary** of a distribution of quantitative data consists of the minimum, the first quartile Q_1, the median, the third quartile Q_3, and the maximum.

Starnes/Tabor, *The Practice of Statistics*

47



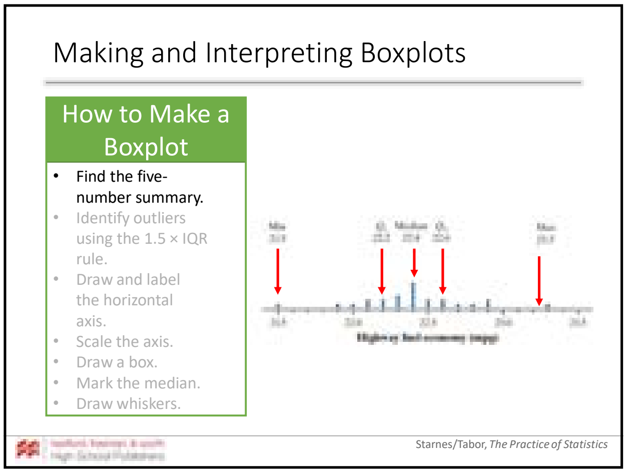
Making and Interpreting Boxplots

The **five-number summary** of a distribution of quantitative data consists of the minimum, the first quartile Q_1, the median, the third quartile Q_3, and the maximum.

A **boxplot** is a visual representation of the five-number summary.

Starnes/Tabor, *The Practice of Statistics*

48



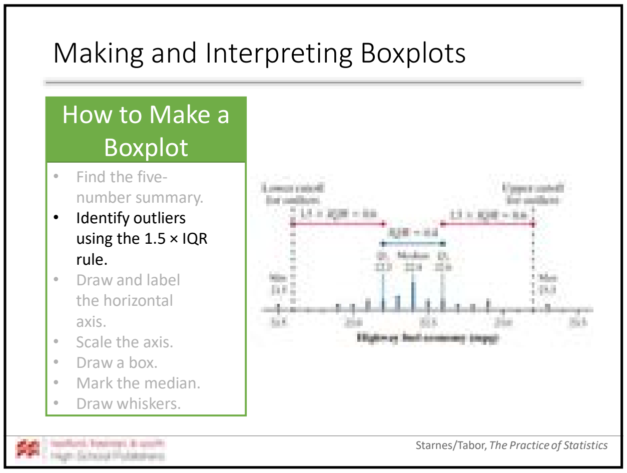
Making and Interpreting Boxplots

How to Make a Boxplot

- Find the five-number summary.
- Identify outliers using the 1.5 × IQR rule.
- Draw and label the horizontal axis.
- Scale the axis.
- Draw a box.
- Mark the median.
- Draw whiskers.

Starnes/Tabor, *The Practice of Statistics*

49



Making and Interpreting Boxplots

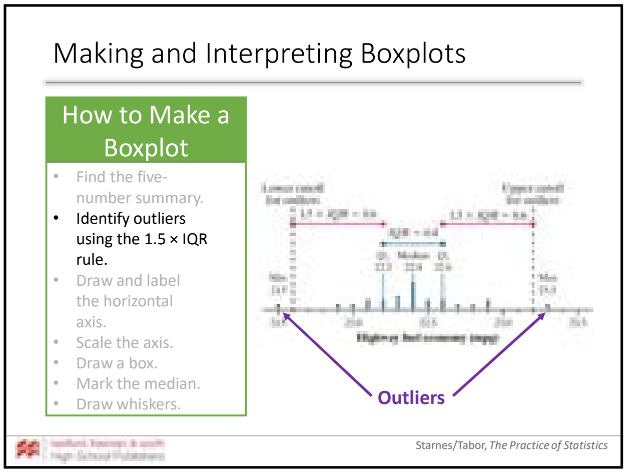
How to Make a Boxplot

- Find the five-number summary.
- Identify outliers using the 1.5 × IQR rule.
- Draw and label the horizontal axis.
- Scale the axis.
- Draw a box.
- Mark the median.
- Draw whiskers.

Lower cutoff for outliers
Upper cutoff for outliers
1.5 × IQR
1.5 × IQR
IQR
Q1
Median
Q3
Min
Max
Highway fuel economy (mpg)

Starnes/Tabor, *The Practice of Statistics*

50



Making and Interpreting Boxplots

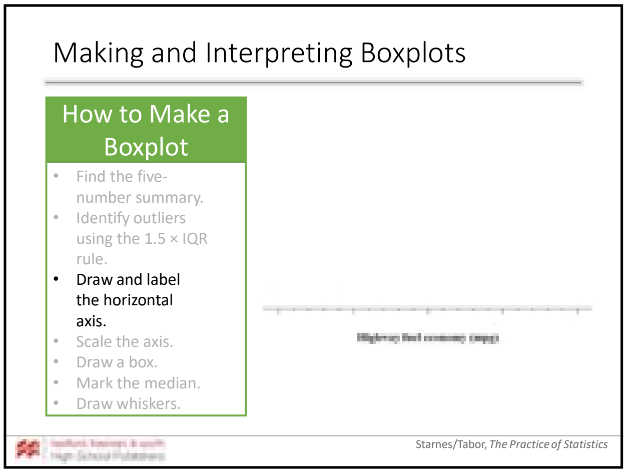
How to Make a Boxplot

- Find the five-number summary.
- Identify outliers using the 1.5 × IQR rule.
- Draw and label the horizontal axis.
- Scale the axis.
- Draw a box.
- Mark the median.
- Draw whiskers.

Lower cutoff for outliers
Upper cutoff for outliers
1.5 × IQR
1.5 × IQR
IQR
Q1
Median
Q3
Min
Max
Highway fuel economy (mpg)
Outliers

Starnes/Tabor, *The Practice of Statistics*

51



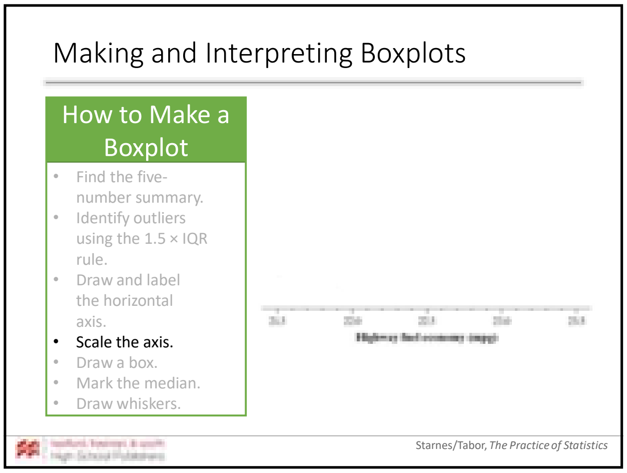
Making and Interpreting Boxplots

How to Make a Boxplot

- Find the five-number summary.
- Identify outliers using the 1.5 × IQR rule.
- Draw and label the horizontal axis.
- Scale the axis.
- Draw a box.
- Mark the median.
- Draw whiskers.

Starnes/Tabor, *The Practice of Statistics*

52



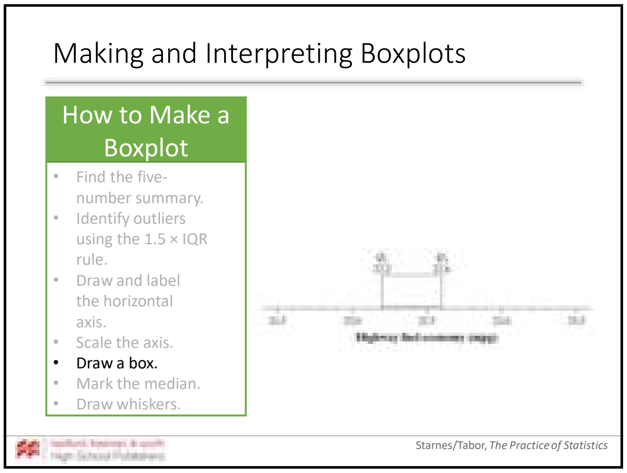
Making and Interpreting Boxplots

How to Make a Boxplot

- Find the five-number summary.
- Identify outliers using the 1.5 × IQR rule.
- Draw and label the horizontal axis.
- Scale the axis.
- Draw a box.
- Mark the median.
- Draw whiskers.

Starnes/Tabor, *The Practice of Statistics*

53



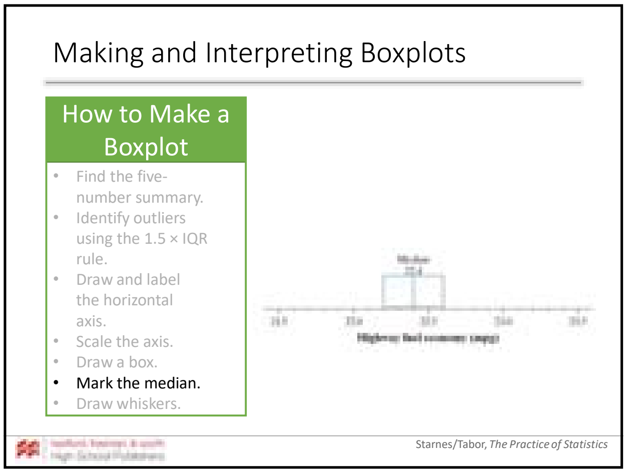
Making and Interpreting Boxplots

How to Make a Boxplot

- Find the five-number summary.
- Identify outliers using the 1.5 × IQR rule.
- Draw and label the horizontal axis.
- Scale the axis.
- Draw a box.
- Mark the median.
- Draw whiskers.

Starnes/Tabor, *The Practice of Statistics*

54



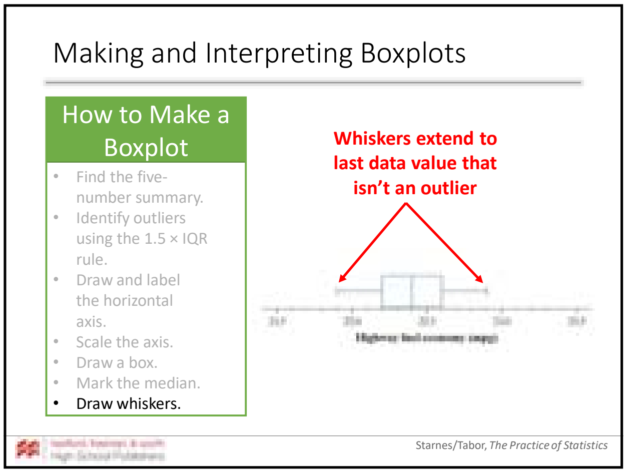
Making and Interpreting Boxplots

How to Make a Boxplot

- Find the five-number summary.
- Identify outliers using the 1.5 × IQR rule.
- Draw and label the horizontal axis.
- Scale the axis.
- Draw a box.
- Mark the median.
- Draw whiskers.

Starnes/Tabor, *The Practice of Statistics*

55



Making and Interpreting Boxplots

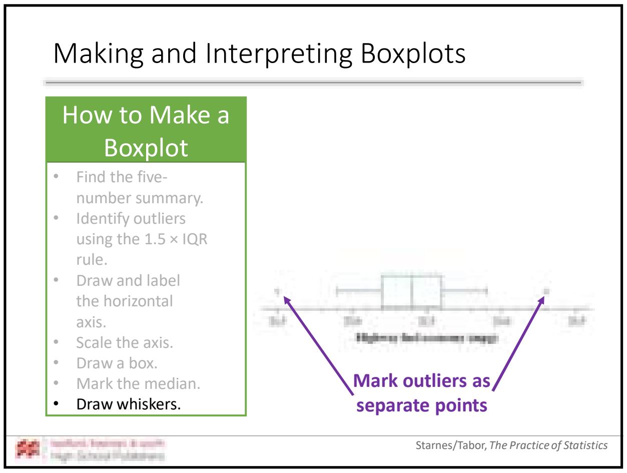
How to Make a Boxplot

- Find the five-number summary.
- Identify outliers using the 1.5 × IQR rule.
- Draw and label the horizontal axis.
- Scale the axis.
- Draw a box.
- Mark the median.
- Draw whiskers.

Whiskers extend to last data value that isn't an outlier

Starnes/Tabor, *The Practice of Statistics*

56



Making and Interpreting Boxplots

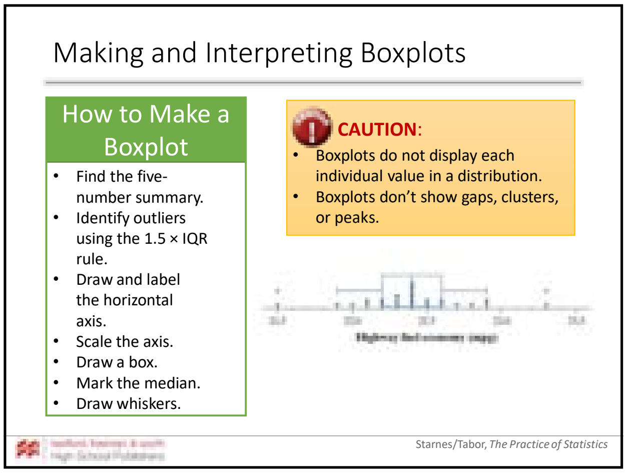
How to Make a Boxplot

- Find the five-number summary.
- Identify outliers using the 1.5 × IQR rule.
- Draw and label the horizontal axis.
- Scale the axis.
- Draw a box.
- Mark the median.
- Draw whiskers.

Mark outliers as separate points

Starnes/Tabor, *The Practice of Statistics*

57



Making and Interpreting Boxplots

How to Make a Boxplot

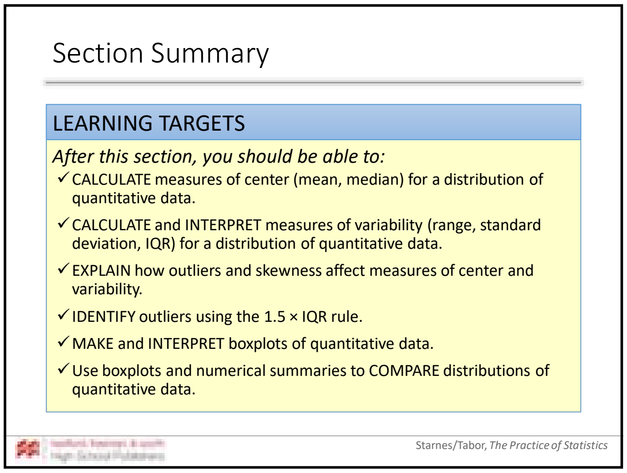
- Find the five-number summary.
- Identify outliers using the 1.5 × IQR rule.
- Draw and label the horizontal axis.
- Scale the axis.
- Draw a box.
- Mark the median.
- Draw whiskers.

CAUTION:

- Boxplots do not display each individual value in a distribution.
- Boxplots don't show gaps, clusters, or peaks.

Starnes/Tabor, *The Practice of Statistics*

58



Section Summary

LEARNING TARGETS

After this section, you should be able to:

✓ CALCULATE measures of center (mean, median) for a distribution of quantitative data.

✓ CALCULATE and INTERPRET measures of variability (range, standard deviation, IQR) for a distribution of quantitative data.

✓ EXPLAIN how outliers and skewness affect measures of center and variability.

✓ IDENTIFY outliers using the 1.5 × IQR rule.

✓ MAKE and INTERPRET boxplots of quantitative data.

✓ Use boxplots and numerical summaries to COMPARE distributions of quantitative data.

Starnes/Tabor, *The Practice of Statistics*

59

1.0 Introduction to Statistics

<u>Read the accompanying slides and answer the following questions</u>

1) What's the difference between categorical and quantitative variables?

2) Do we ever use numbers to describe the values of a categorical variable? Do we ever divide the distribution of a quantitative variable into categories?

Here is information about 8 randomly selected US residents from the 2000 census.

State	Number of family members	Age	Marital status	Travel time to work
Kentucky	2	61	Married	20
Florida	6	27	Married	20
Michigan	3	49	Married	25
Virginia	3	26	Married	15
Pennsylvania	4	44	Married	10
Virginia	4	22	Never married/ single	0
California	1	30	Never married/ single	15
New York	4	34	Separated	40

3) Who are the individuals in this data set?

4) What variables are measured? Identify each as categorical or quantitative.

5) For quantitative variables, what is the difference between a discrete and a continuous variable?

1.1 Analyzing Categorical Data

6) What is the difference between a data table, a frequency table, and a relative frequency table? When is it better to use relative frequency?

7) What is the most important thing to remember when making pie charts and bar graphs? Why do statisticians prefer bar graphs?

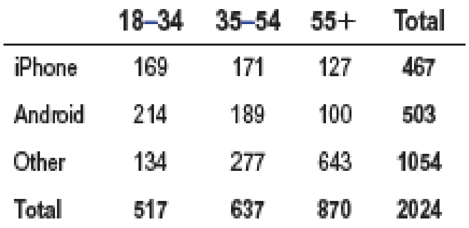
8) What are some common ways to make a misleading graph?

9) What is a two-way table? What is a marginal relative frequency?

10) What is a joint relative frequency?

11) What is a conditional relative frequency?

The Pew Research Center asked a random sample of 2024 adult cell phone owners from the United States which type of cell phone they own: iPhone, Android, or other (including non-smart phones). Here are the results, broken down by age category:



	18–34	35–54	55+	Total
iPhone	169	171	127	**467**
Android	214	189	100	**503**
Other	134	277	643	**1054**
Total	**517**	**637**	**870**	**2024**

12) What proportion of the sample use an iPhone?

13) What proportion of the sample use an iPhone and are 55+?

14) What proportion of the 55+ people in the sample use an iPhone?

15) What proportion of the iPhone users in the sample are 55+?

16) What does it mean for two variables to have an association?

17) How can you “see” an association between two categorical variables?

18) Explain what it would mean if there was no association between age and cell phone type.

19) Display the relationship between age group and cell phone type using a mosaic plot. Based on the graph, is there an association between age and cell phone type? Justify.

1.2 Displaying Quantitative Data with Graphs

Overall pattern of a distribution

1) What is a distribution?

When DESCRIBING ALL DISTRIBUTIONS, you must include the following: **SOCV** **[VERY IMPORTANT]**

SHAPE:

2) Briefly illustrate the following distribution shapes:

Symmetric	Skewed right	Skewed left

Unimodal (Single-peaked)	Bimodal (Double-peaked)	Uniform (no peaks)

OUTLIERS:

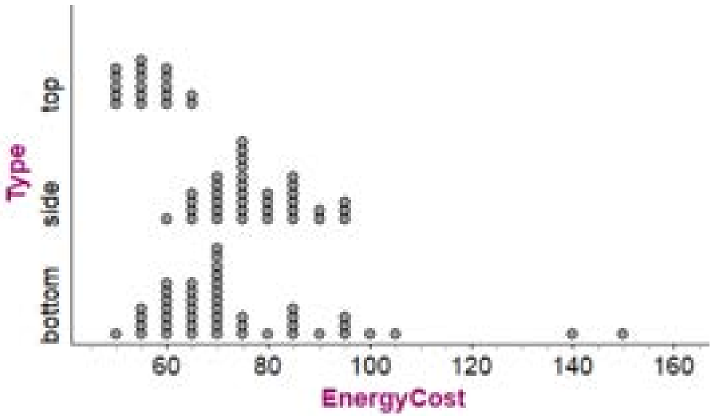
CENTER:

VARIABILITY:

3) How do you describe a distribution of a quantitative variable?

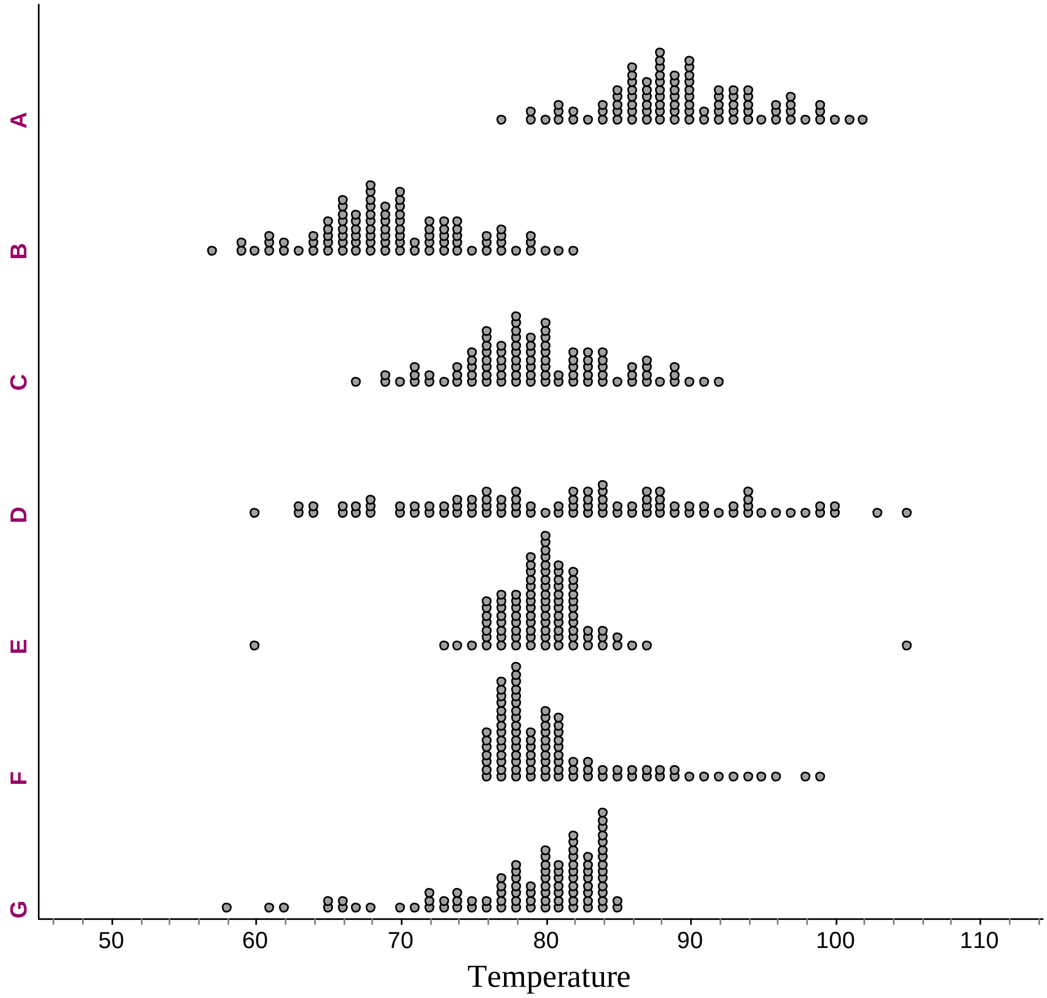
4) What are the 2 most important things to remember when you are asked to compare distributions?

5) How do the annual energy costs (in dollars) compare for refrigerators with top freezers, side freezers, and bottom freezers? The data below is from the May 2010 issue of *Consumer Reports*. **Compare these distributions**.



Dotplots

Brian and Jessica have decided to move and are considering seven different cities. The dotplots below show the daily high temperatures in June, July, and August for each of these cities. Help them pick a city by answering the questions below.



1) What is the most important difference between cities A, B, and C?

2) What is the most important difference between cities C and D?

3) What are two important differences between cities D and E?

4) What is the most important difference between cities C, F, and G?

Stemplots

1) What is the most important thing to remember when making a stemplot?

2) A sample of 14-year-olds from the United Kingdom was randomly selected. Here are the heights of the students (in cm). **Make a back-to-back stemplot and compare the distributions.**

Male: 154, 157, 187, 163, 167, 159, 169, 162, 176, 177, 151, 175, 174, 165, 165, 183, 180

Female: 160, 169, 152, 167, 164, 163, 160, 163, 169, 157, 158, 153, 161, 165, 165, 159, 168,
153, 166, 158, 158, 166

Histograms

1) How do you make a histogram?

2) How is a histogram different than a bar chart?

3) Why would we prefer a *relative* frequency histogram to a frequency histogram?

4) What will cause you to lose points on tests and projects (and make Mr. Denny lose years from his life)?

The following table presents the average points scored per game (PPG) for the 30 NBA teams in a recent season. **Make a dotplot to display the distribution of points per game. Then, make a histogram**.

Team	**PPG**	**Team**	**PPG**	**Team**	**PPG**
Atlanta Hawks	98.0	Houston Rockets	106.0	Oklahoma City Thunder	105.7
Boston Celtics	96.5	Indiana Pacers	94.7	Orlando Magic	94.1
Brooklyn Nets	96.9	Los Angeles Clippers	101.1	Philadelphia 76ers	93.2
Charlotte Bobcats	93.4	Los Angeles Lakers	102.2	Phoenix Suns	95.2
Chicago Bulls	93.2	Memphis Grizzlies	93.4	Portland Trail Blazers	97.5
Cleveland Cavaliers	96.5	Miami Heat	102.9	Sacramento Kings	100.2
Dallas Mavericks	101.1	Milwaukee Bucks	98.9	San Antonio Spurs	103.0
Denver Nuggets	106.1	Minnesota Timberwolves	95.7	Toronto Raptors	97.2
Detroit Pistons	94.9	New Orleans Hornets	94.1	Utah Jazz	98.0
Golden State Warriors	101.2	New York Knicks	100.0	Washington Wizards	93.2

Dotplot

Histogram

1.3 Describing Quantitative Data with Numbers

Measuring Center (Median and Mean)

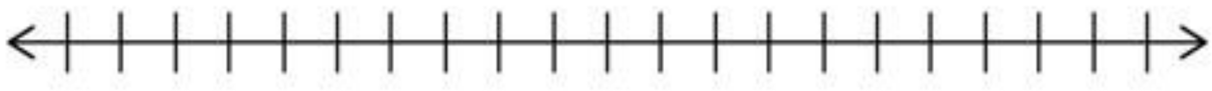
5) What is the difference between a statistic and a parameter?

6) The following data are travel times for fifteen people to get to work in minutes:

20 30 10 40 25 20 10 60 15 40 5 30 12 10 10

Rewrite the numbers in order from least to greatest:

Make a dotplot of the data for a visual representation:



7) Define **Median**: (Both via words and mathematically) and what is the median of the data set above.

8) Define **Mean**: (Both via words and mathematically) and what is the mean of the data set above.

Comparing the Mean and Median:
The mean and median of a roughly symmetric distribution are close together. If the distribution is exactly symmetric, the mean and median are the same. In a skewed distribution the mean is usually farther out in the long tail than its median. If the outliers were to increase, it would increase the mean, but the median would stay the same.

9) What is a resistant measure? Is the mean a resistant measure of center?

10) How can you estimate the mean of a histogram or dotplot?

11) Is the median a resistant measure of center? Explain.

12) How do skewness and outliers affect the relationship between the mean and the median?

Measuring Variability

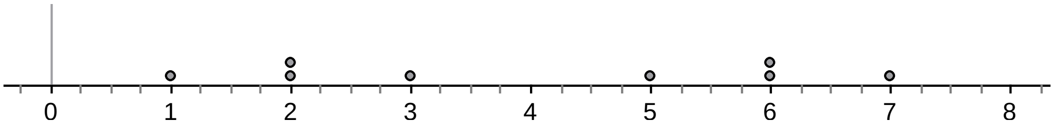
Range

1. What is the range? How is it calculated mathematically? What is the range of the data set above (Driving Times)?

2. What are two problems with range as a measure of variability?

Standard Deviation

In the distribution below, how far are the values from the mean, on average?



Define **Standard Deviation** & what does the standard deviation measure?

How do you calculate the standard deviation for a population? What about the variance?

How do you calculate the standard deviation for a sample?

What are some properties of the standard deviation?

A random sample of 5 students was asked how many minutes they spent doing HW the previous night. Here are their responses (in minutes): 0, 25, 30, 60, 90. **Calculate *and* interpret the standard deviation**.

The Interquartile Range

1) What are quartiles? How do you find them?

2) What is the **interquartile range (*IQR*)**? Is the *IQR* a resistant measure of variability?

The table shows the number of runs the Cubs allowed to score during day games in two different types of weather. For each distribution, **calculate the *IQR***.

Cloudy:	0	1	1	2	3	3	3	3	3	3	4	4	4	4	4	4	5	6	6	6	6	9	9	10	11	13	14

Sunny:	0	0	1	2	2	2	3	3	3	4	5	5	5	5	8	11	12	15

3) How do you calculate summary statistics using the calculator?

Identifying Outliers

1) What is an outlier? How do you identify them? Check out the *IQR* dance. https://youtu.be/mfX7l--CIs4?si=wjrIyOJ6WOzVNcBf

2) Are there any outliers in the runs allowed distributions from the data in the notes above? Justify.

3) What is the **five-number summary**? How is it displayed?

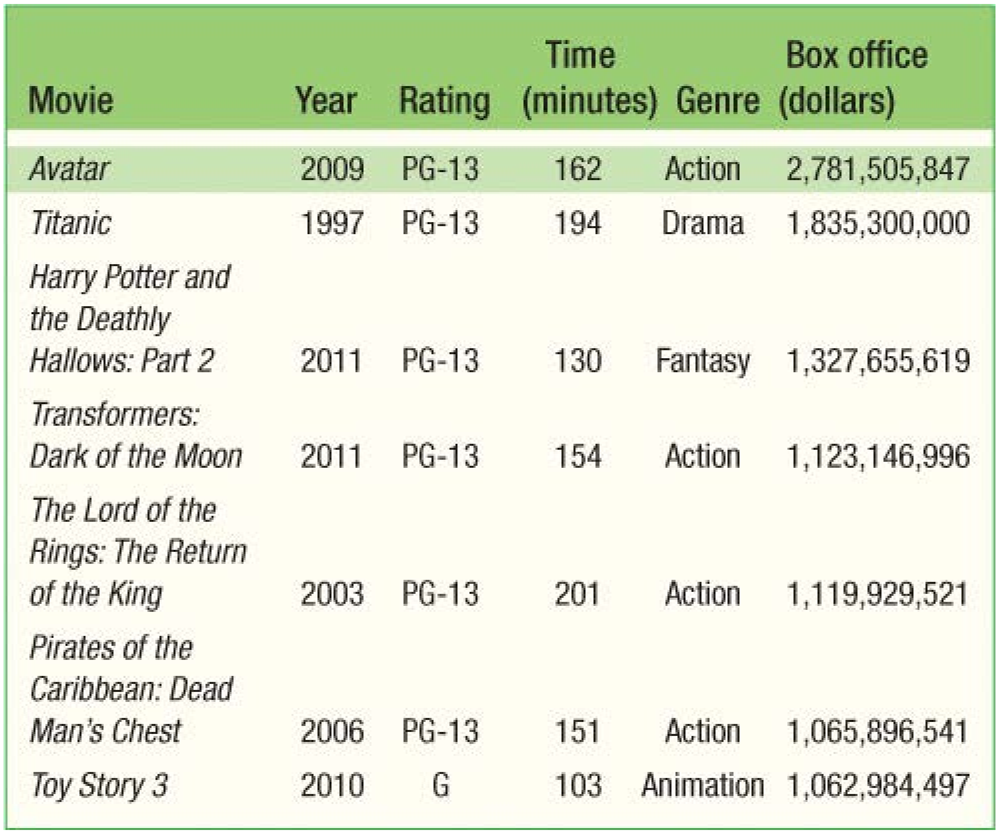
4) Draw parallel boxplots for Cubs cloudy/sunny data. Compare these distributions.

5) What are some weaknesses of boxplots?

Chapter 1 Chapter Review Exercises

These exercises are designed to help you review the important ideas and methods of the chapter.

R1.1. Hit movies According to the Internet Movie Database, *Avatar* is tops based on box office sales worldwide. The following table displays data on several popular movies.[47]



			Time		Box office
Movie	Year	Rating	(minutes)	Genre	(dollars)
Avatar	2009	PG-13	162	Action	2,781,505,847
Titanic	1997	PG-13	194	Drama	1,835,300,000
Harry Potter and the Deathly Hallows: Part 2	2011	PG-13	130	Fantasy	1,327,655,619
Transformers: Dark of the Moon	2011	PG-13	154	Action	1,123,146,996
The Lord of the Rings: The Return of the King	2003	PG-13	201	Action	1,119,929,521
Pirates of the Caribbean: Dead Man's Chest	2006	PG-13	151	Action	1,065,896,541
Toy Story 3	2010	G	103	Animation	1,062,984,497

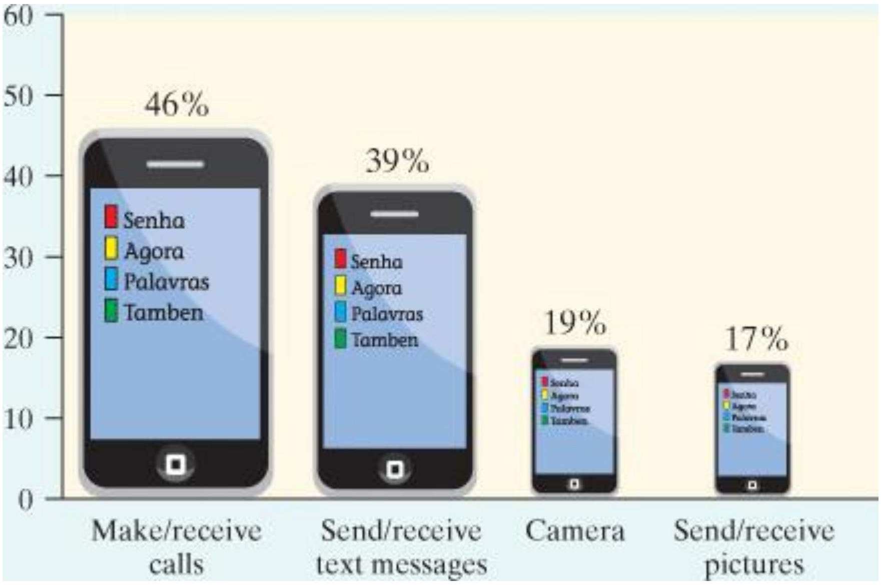
(a) What individuals does this data set describe?

(b) Clearly identify each of the variables. Which are quantitative?

(c) Describe the individual in the highlighted row.

R1.2. Movie ratings The movie rating system we use today was first established on November 1, 1968. Back then, the possible ratings were G, PG, R, and X. In 1984, the PG-13 rating was created. And in 1990, NC-17 replaced the X rating. Here is a summary of the ratings assigned to movies between 1968 and 2000: 8% rated G, 24% rated PG,10% rated PG-13, 55% rated R, and 3% rated NC-17.[48] Make an appropriate graph for displaying these data.

R1.3. I'd die without my phone! In a survey of over 2000 U.S. teenagers by Harris Interactive, 47% said that "their social life would end or be worsened without their cell phone."[49] One survey question asked the teens how important it is for their phone to have certain features. The figure below displays data on the percent who indicated that a particular feature is vital.

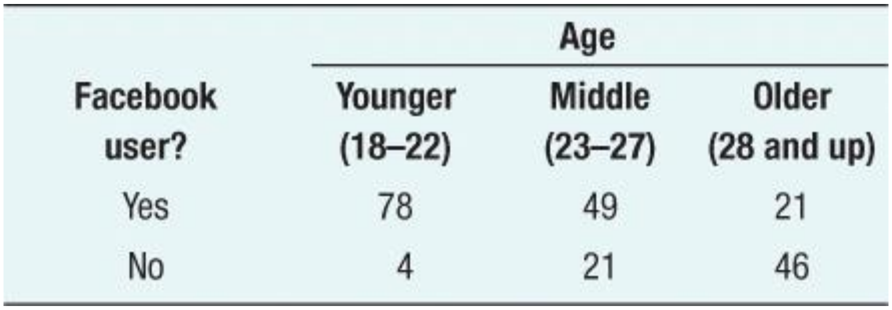


(a) Explain how the graph gives a misleading impression.

(b) Would it be appropriate to make a pie chart to display these data? Why or why not?

(c) Make a graph of the data that isn't misleading.

R1.4. Facebook and age Is there a relationship between Facebook use and age among college students? The following two-way table displays data for the 219 students who responded to the survey.[50]



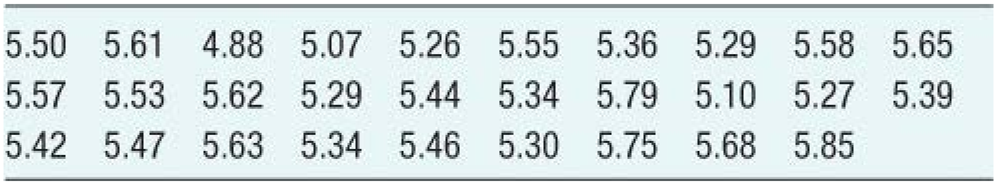
	Age		
Facebook user?	Younger (18–22)	Middle (23–27)	Older (28 and up)
Yes	78	49	21
No	4	21	46

(a) What percent of the students who responded were Facebook users? Is this percent part of a marginal distribution or a conditional distribution? Explain.

(b) What percent of the younger students in the sample were Facebook users? What percent of the Facebook users in the sample were younger students?

R1.5. Facebook and age Use the data in the previous exercise to determine whether there is an association between Facebook use and age. Give appropriate graphical and numerical evidence to support your answer.

R1.6. Density of the earth In 1798, the English scientist Henry Cavendish measured the density of the earth several times by careful work with a torsion balance. The variable recorded was the density of the earth as a multiple of the density of water. Here are Cavendish's 29 measurements:[51]



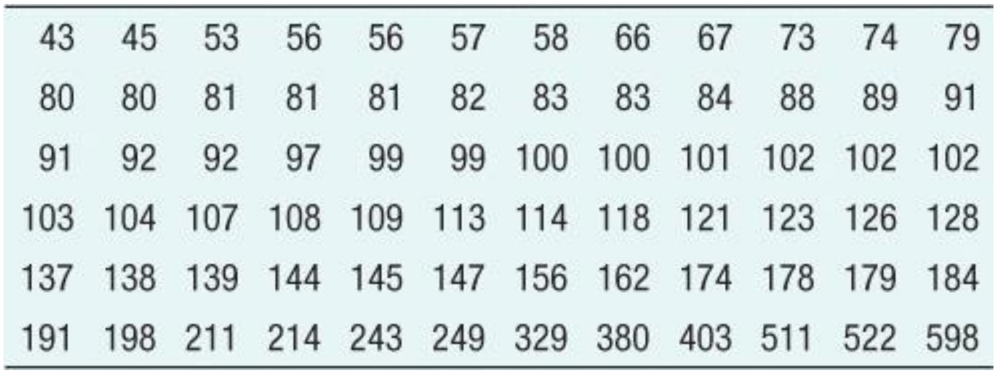
5.50	5.61	4.88	5.07	5.26	5.55	5.36	5.29	5.58	5.65
5.57	5.53	5.62	5.29	5.44	5.34	5.79	5.10	5.27	5.39
5.42	5.47	5.63	5.34	5.46	5.30	5.75	5.68	5.85	

(a) Present these measurements graphically in a stemplot.

(b) Discuss the shape, center, and spread of the distribution. Are there any outliers?

(c) What is your estimate of the density of the earth based on these measurements? Explain.

R1.7. Guinea pig survival times Here are the survival times in days of 72 guinea pigs after they were injected with infectious bacteria in a medical experiment.[52] Survival times, whether of machines under stress or cancer patients after treatment, usually have distributions that are skewed to the right.



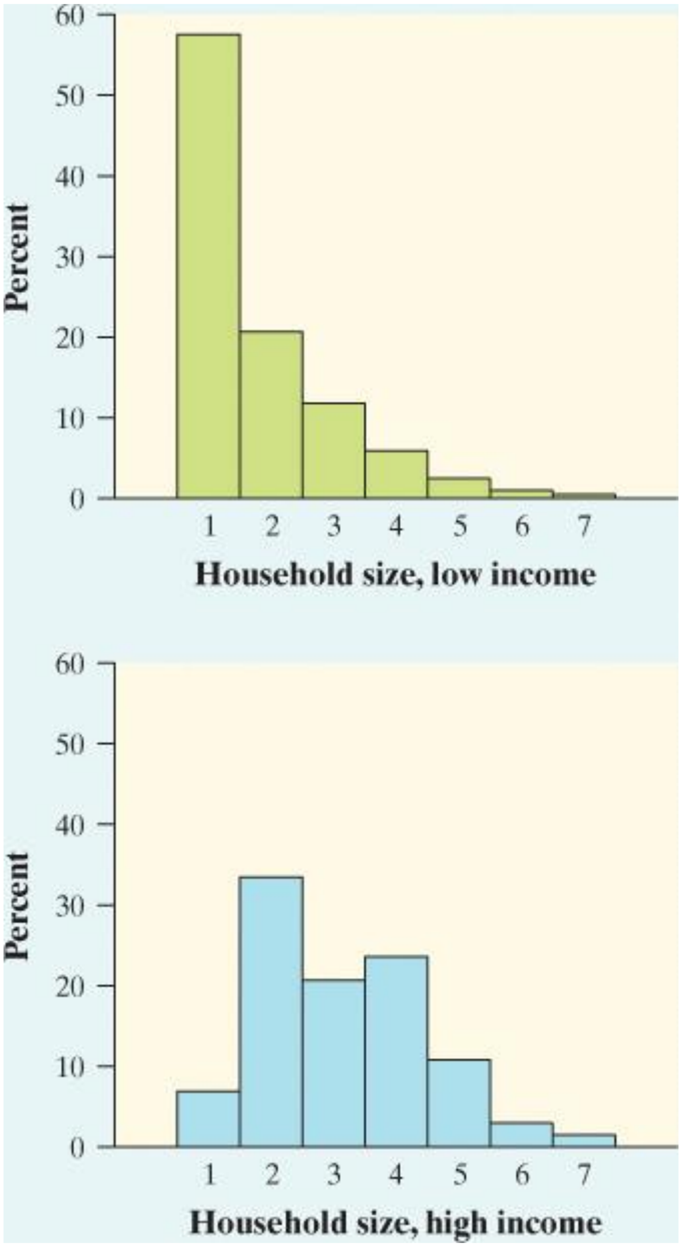
43	45	53	56	56	57	58	66	67	73	74	79
80	80	81	81	81	82	83	83	84	88	89	91
91	92	92	97	99	99	100	100	101	102	102	102
103	104	107	108	109	113	114	118	121	123	126	128
137	138	139	144	145	147	156	162	174	178	179	184
191	198	211	214	243	249	329	380	403	511	522	598

(a) Make a histogram of the data and describe its main features. Does it show the expected right skew?

(b) Now make a boxplot of the data. Be sure to check for outliers.

(c) Which measure of center and spread would you use to summarize the distribution—the mean and standard deviation or the median and *IQR?* Justify your answer.

R1.8. Household incomes Rich and poor households differ in ways that go beyond income. Following are histograms that compare the distributions of household size(number of people) for low-income and high-income households.[53] Low-income households had annual incomes less than $15,000, and high-income households had annual incomes of at least $100,000.



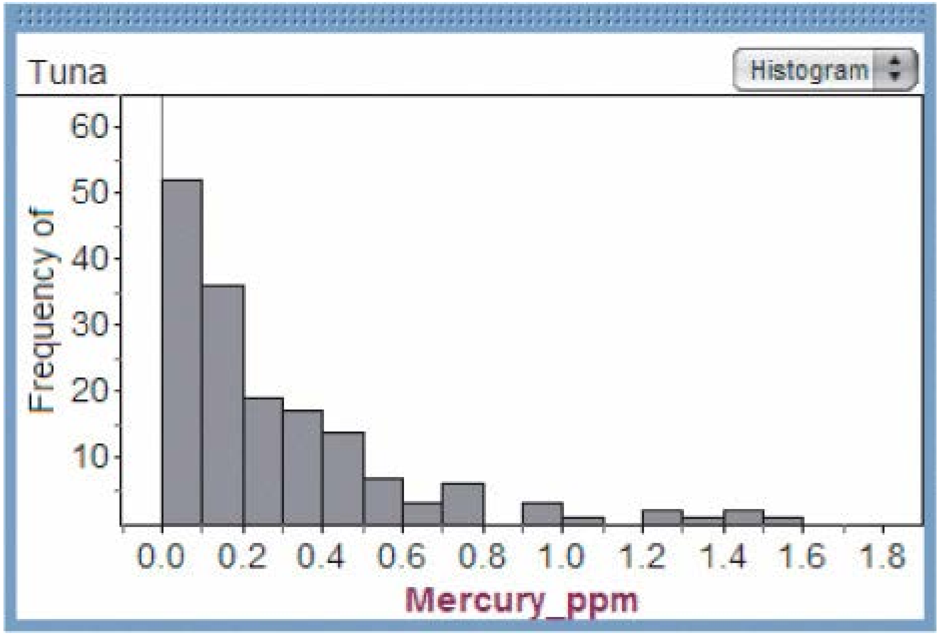
(a) About what percent of each group of households consisted of two people?

(b) What are the important differences between these two distributions? What do you think explains these differences?

Exercises R1.9 and R1.10 refer to the following setting. Do you like to eat tuna? Many people do. Unfortunately, some of the tuna that people eat may contain high levels of mercury. Exposure to mercury can be especially hazardous for pregnant women and small children. How much mercury is safe to consume? The Food and Drug Administration will take action (like removing the product from store shelves) if the mercury concentration in a six-ounce can of tuna is 1.00 ppm (parts per million) or higher.

What is the typical mercury concentration in cans of tuna sold in stores? A study conducted by Defenders of Wildlife set out to answer this question. Defenders collected a sample of 164 cans of tuna from stores across the United States. They sent the selected cans to a laboratory that is often used by the Environmental Protection Agency for mercury testing.[54]

R1.9. Mercury in tuna A histogram and some computer output provide information about the mercury concentration in the sampled cans (in parts per million, ppm).



Descriptive Statistics: Mercury_ppm

Variable	N	Mean	StDev	Min
Mercury	164	0.285	0.300	0.012

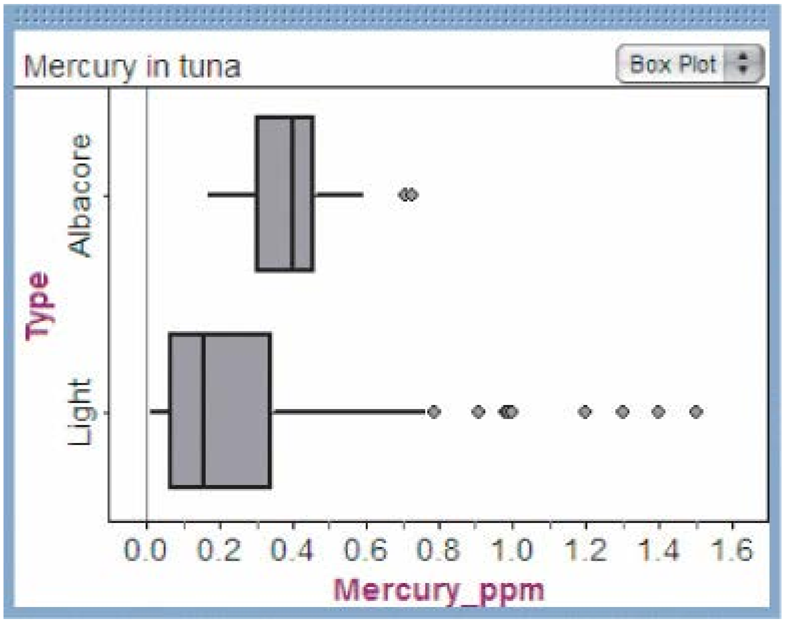
Variable	Q_1	Med	Q_3	Max
Mercury	0.071	0.180	0.380	1.500

(a) Interpret the standard deviation in context.

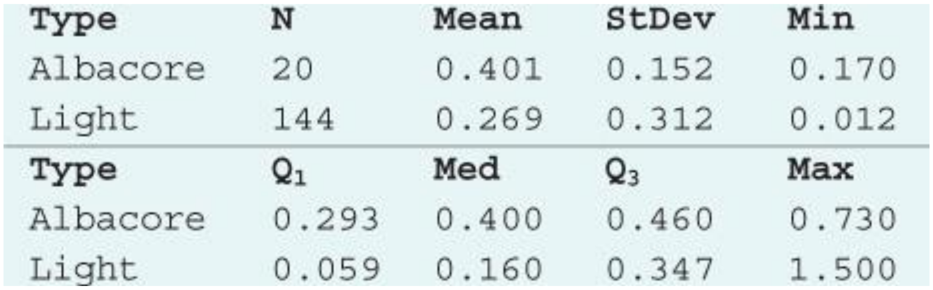
(b) Determine whether there are any outliers.

(c) Describe the shape, center, and spread of the distribution.

R1.10. Mercury **in tuna** Is there a difference in the mercury concentration of light tuna and albacore tuna? Use the parallel boxplots and the computer output to write a few sentences comparing the two distributions.



Descriptive Statistics: Mercury_ppm



Type	N	Mean	StDev	Min
Albacore	20	0.401	0.152	0.170
Light	144	0.269	0.312	0.012
Type	**Q_1**	**Med**	**Q_3**	**Max**
Albacore	0.293	0.400	0.460	0.730
Light	0.059	0.160	0.347	1.500

1.6 Chapter 1: AP® Statistics Practice Test

Section I: Multiple Choice *Select the best answer for each question*.

T1.1. You record the age, marital status, and earned income of a sample of 1463 women. The number and type of variables you have recorded is

(a) 3 quantitative, 0 categorical.

(b) 4 quantitative, 0 categorical.

(c) 3 quantitative, 1 categorical.

(d) 2 quantitative, 1 categorical.

(e) 2 quantitative, 2 categorical.

T1.2. Consumers Union measured the gas mileage in miles per gallon of 38 vehicles from the same model year on a special test track. The pie chart provides information about the country of manufacture of the model cars tested by Consumers Union. Based on the pie chart, we conclude that

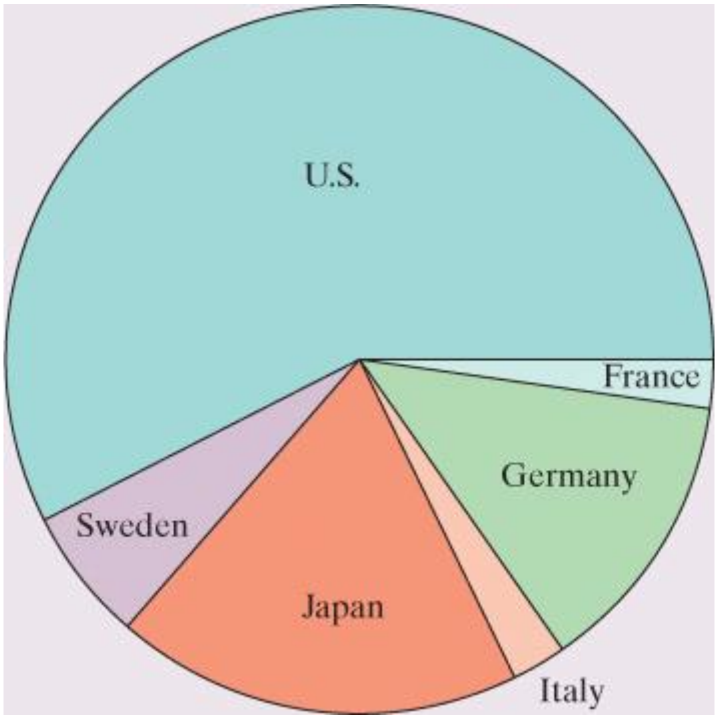
(a) Japanese cars get significantly lower gas mileage than cars from other countries.

(b) U.S. cars get significantly higher gas mileage than cars from other countries.

(c) Swedish cars get gas mileages that are between those of Japanese and U.S. cars.

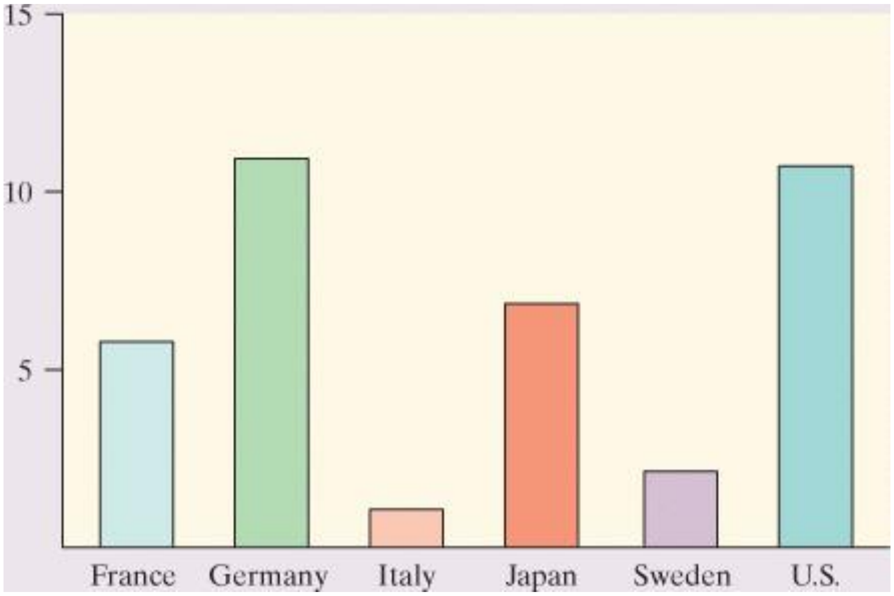
(d) cars from France have the lowest gas mileage.

(e) more than half of the cars in the study were from the United States.

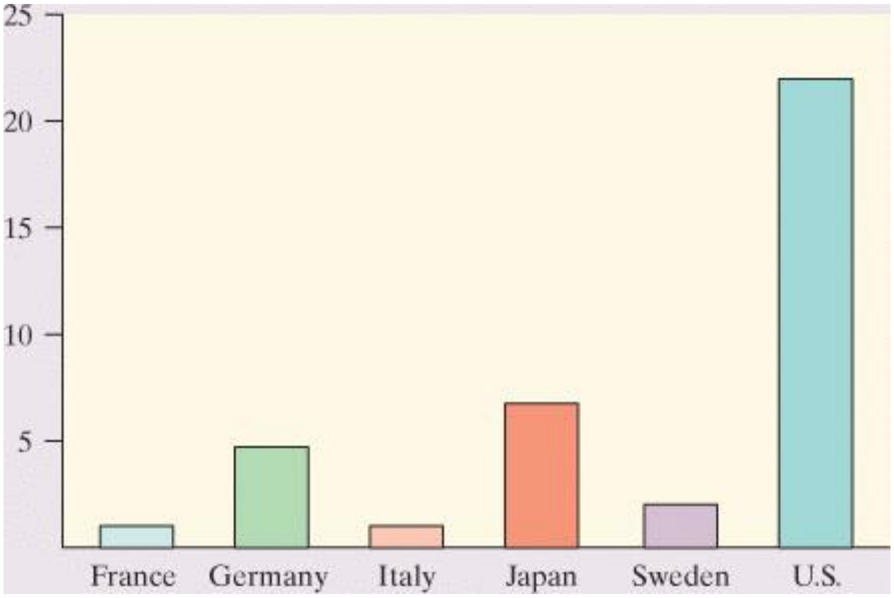


T1.3. Which of the following bar graphs is equivalent to the pie chart in Question T1.2?

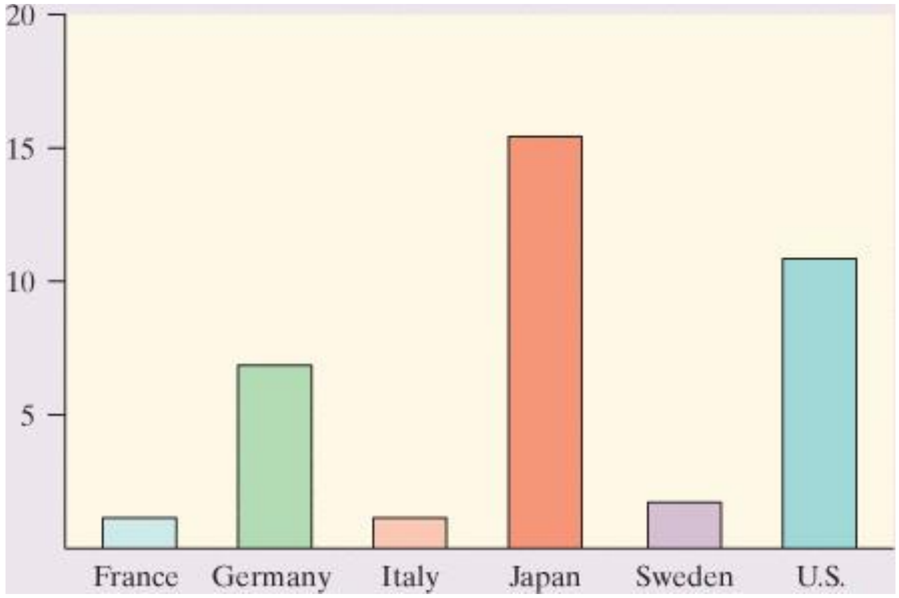
(a)


15
10
5
France
Germany
Italy
Japan
Sweden
U.S.

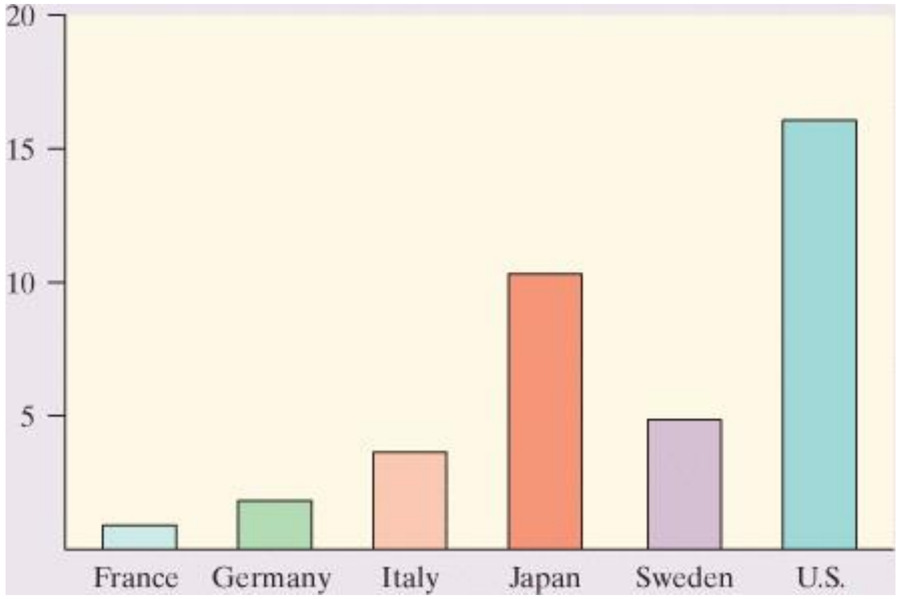
(b)


25
20
15
10
5
France
Germany
Italy
Japan
Sweden
U.S.

(c)



(d)



(e) None of these.

T1.4. Earthquake intensities are measured using a device called a seismograph, which is designed to be most sensitive to earthquakes with intensities between 4.0 and 9.0 on the Richter scale. Measurements of nine earthquakes gave the following readings:

4.5	L	5.5	H	8.7	8.9	6.0	H	5.2

where L indicates that the earthquake had an intensity below 4.0 and an H indicates that the earthquake had an intensity above 9.0. The median earthquake intensity of the sample is

(a) 5.75.

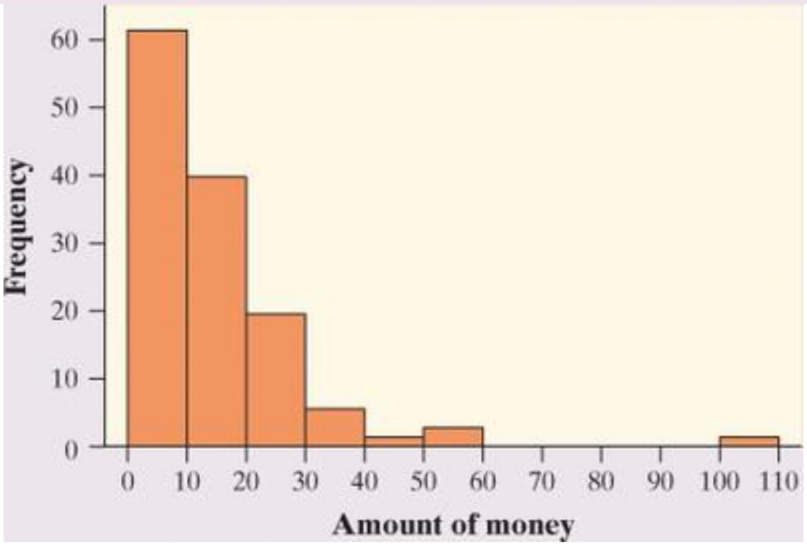
(b) 6.00.

(c) 6.47.

(d) 8.70.

(e) Cannot be determined.

*Questions T*1.5 *and T*1.6 *refer to the following setting*. In a statistics class with 136 students, the professor records how much money (in dollars) each student has in his or her possession during the first class of the semester. The histogram shows the data that were collected.



T1.5. The percentage of students with less than $10 in their possession is closest to

(a) 30%.

(b) 35%.

(c) 45%.

(d) 60%.

(e) 70%.

T1.6. Which of the following statements about this distribution is *not* correct?

(a) The histogram is right-skewed.

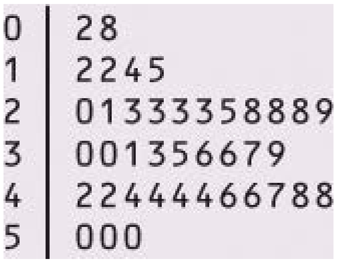
(b) The median is less than $20.

(c) The *IQR* is $35.

(d) The mean is greater than the median.

(e) The histogram is unimodal.

T1.7. Forty students took a statistics examination having a maximum of 50 points. The score distribution is given in the following stem-and-leaf plot:



0	28
1	2245
2	01333358889
3	001356679
4	22444466788
5	000

The third quartile of the score distribution is equal to

(a) 45.

(b) 44.

(c) 43.

(d) 32.

(e) 23.

T1.8. The mean salary of all female workers is $35,000. The mean salary of all male workers is $41,000. What must be true about the mean salary of all workers?

(a) It must be $38,000.

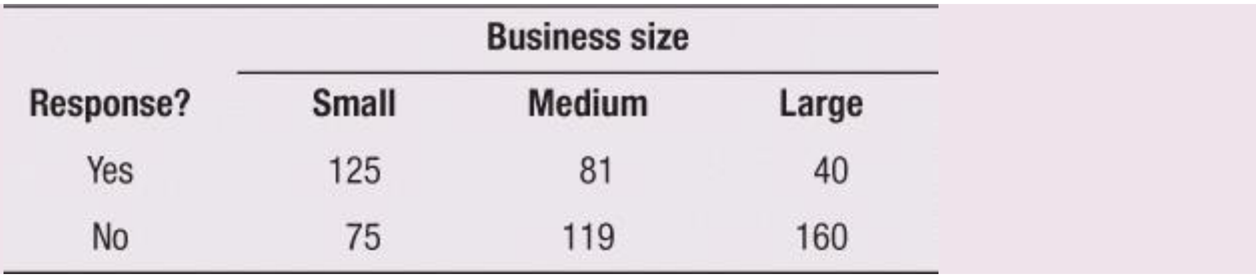
(b) It must be larger than the median salary.

(c) It could be any number between $35,000 and $41,000.

(d) It must be larger than $38,000.

(e) It cannot be larger than $40,000.

Questions T1.9 and T1.10 refer to the following setting. A survey was designed to study how business operations vary according to their size. Companies were classified as small, medium, or large. Questionnaires were sent to 200 randomly selected businesses of each size. Because not all questionnaires in a survey of this type are returned, researchers decided to investigate the relationship between the response rate and the size of the business. The data are given in the following two-way table:



	Business size		
Response?	Small	Medium	Large
Yes	125	81	40
No	75	119	160

T1.9. What percent of all small companies receiving questionnaires responded?

(a) 12.5%

(b) 20.8%

(c) 33.3%

(d) 50.8%

(e) 62.5%

T1.10. Which of the following conclusions seems to be supported by the data?

(a) There are more small companies than large companies in the survey.

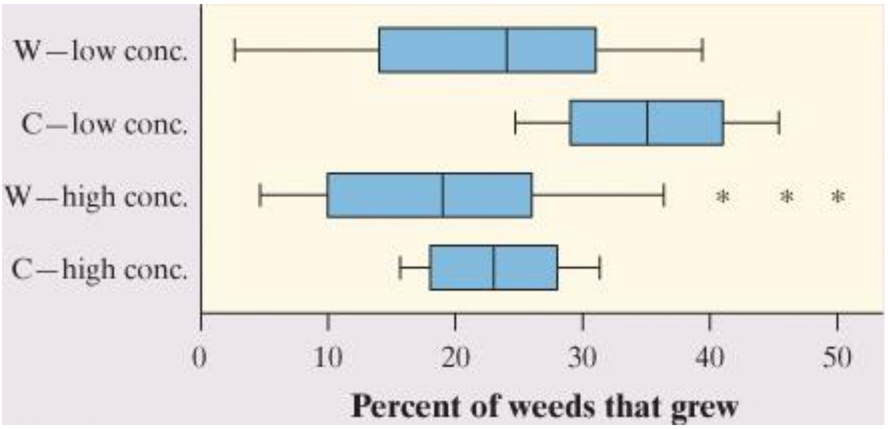
(b) Small companies appear to have a higher response rate than medium or big companies.

(c) Exactly the same number of companies responded as didn't respond.

(d) Overall, more than half of companies responded to the survey.

(e) If we combined the medium and large companies, then their response rate would be equal to that of the small companies.

T1.11. An experiment was conducted to investigate the effect of a new weed killer to prevent weed growth in onion crops. Two chemicals were used: the standard weed killer(C) and the new chemical (W). Both chemicals were tested at high and low concentrations on a total of 50 test plots. The percent of weeds that grew in each plot was recorded. Here are some boxplots of the results. Which of the following is *not* a correct statement about the results of this experiment?



(a) At both high and low concentrations, the new chemical (W) gives better weed control than the standard weed killer (C).

(b) Fewer weeds grew at higher concentrations of both chemicals.

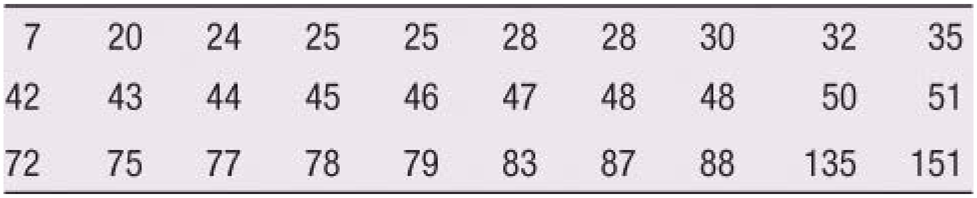
(c) The results for the standard weed killer (C) are less variable than those for the new chemical (W).

(d) High and low concentrations of either chemical have approximately the same effects on weed growth.

(e) Some of the results for the low concentration of weed killer W show fewer weeds growing than some of the results for the high concentration of W.

Section II: Free Response *Show all your work. Indicate clearly the methods you use, because you will be graded on the correctness of your methods as well as on the accuracy and completeness of your results and explanations*.

T1.12. You are interested in how much time students spend on the Internet each day.Here are data on the time spent on the Internet (in minutes) for a particular day reported by a random sample of 30 students at a large high school:



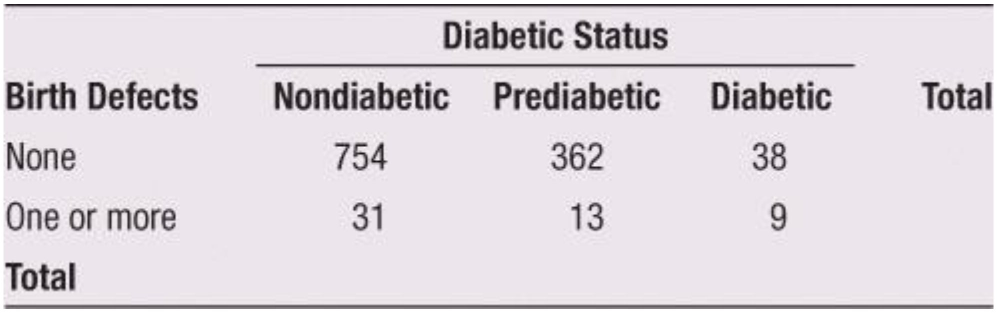
7	20	24	25	25	28	28	30	32	35
42	43	44	45	46	47	48	48	50	51
72	75	77	78	79	83	87	88	135	151

(a) Construct a histogram of these data.

(b) Are there any outliers? Justify your answer.

(c) Would it be better to use the mean and standard deviation or the median and *IQR* to describe the center and spread of this distribution? Why?

T1.13. A study among the Pima Indians of Arizona investigated the relationship between a mother's diabetic status and the appearance of birth defects in her children. The results appear in the two-way table below.



	Diabetic Status			
Birth Defects	**Nondiabetic**	**Prediabetic**	**Diabetic**	**Total**
None	754	362	38	
One or more	31	13	9	
Total				

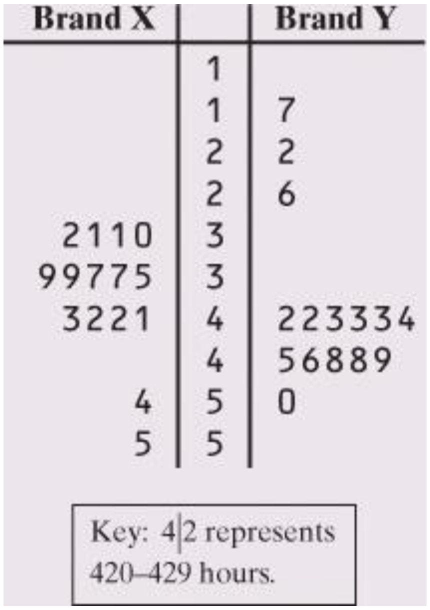
(a) Fill in the row and column totals in the margins of the table.

(b) Compute (in percents) the conditional distributions of birth defects for each diabetic status.

(c) Display the conditional distributions in a graph. Don't forget to label your graph completely.

(d) Do these data give evidence of an association between diabetic status and birth defects? Justify your answer.

T1.14. The back-to-back stemplot shows the lifetimes of several Brand X and Brand Y batteries.



Brand X		Brand Y
	1	
	1	7
	2	2
	2	6
2110	3	
99775	3	
3221	4	223334
	4	56889
4	5	0
5	5	

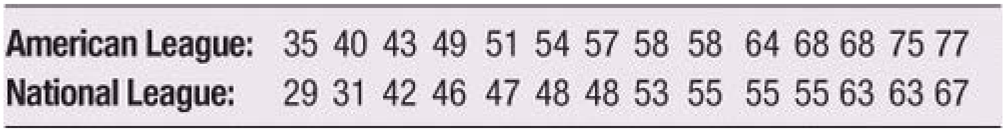
Key: 4|2 represents 420–429 hours.

(a) What is the longest that any battery lasted?

(b) Give a reason someone might prefer a Brand X battery.

(c) Give a reason someone might prefer a Brand Y battery.

T1.15. During the early part of the 1994 baseball season, many fans and players noticed that the number of home runs being hit seemed unusually large. Here are the data on the number of home runs hit by American League and National League teams in the early part of the 1994 season:



American League:	35	40	43	49	51	54	57	58	58	64	68	68	75	77
National League:	29	31	42	46	47	48	48	53	55	55	55	63	63	67

Compare the distributions of home runs for the two leagues graphically and numerically. Write a few sentences summarizing your findings.