Warm Up

1. Statistics is the study of ___**data**___. When referring to a population, a characteristic we reference is called a ___**parameter**___, but when referring to a sample, a characteristic we reference is called a ___***statistic***___.
2. What is a frequency distribution table?
3. What is a *relative frequency* distribution?

Chapter 3: Graphical Methods for Describing Data

Section 3.1

Analyzing Categorical Data

Warm Up

1) A ______________________________ (SRS) of size *n* is chosen in such a way that every group of *n* individuals in the population has an equal _____________ to be selected as the sample.

2) What is a frequency distribution table?

3) What is a *relative frequency* distribution?

Warm Up

1. A ______**simple random sample**______ (SRS) of size *n* is chosen in such a way that every group of *n* individuals in the population has an equal ___**probability**___ to be selected as the sample.
2. What is a frequency distribution table?
3. What is a *relative frequency* distribution?

Frequency Distributions & Bar Charts for Categorical Data

- **Frequency Distribution**: A table that displays the possible categories along with the associated *frequencies* (the count or number of times it occurs)

- **Relative Frequency Distribution**: A table that displays the possible categories along with the proportion of observations for each category.

- $relative\ frequency = \dfrac{frequency}{total\ observations\ in\ data\ set}$

Chapter 3: Graphical Methods for Describing Data

Section 3.1

Analyzing Categorical Data

Chapter 3: Graphical Methods for Describing Data

- **Introduction:** Data Analysis: Making Sense of Data
- **3.1 Review Analyzing Categorical Data**
- **3.2** Displaying Quantitative Data with Graphs
- **3.3** Describing Quantitative Data with Numbers

Section 3.1
Analyzing Categorical Data

Learning Objectives

After this section, you should be able to…

- ✓ CONSTRUCT and INTERPRET bar graphs and pie charts
- ✓ RECOGNIZE “good” and “bad” graphs
- ✓ CONSTRUCT and INTERPRET two-way tables
- ✓ DESCRIBE relationships between two categorical variables
- ✓ ORGANIZE statistical problems

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

Example

Variable

Values

Frequency Table

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

Count

Relative Frequency Table

Format	Percent of Stations
Adult Contemporary	11.2
Adult Standards	8.6
Contemporary Hit	4.1
Country	14.9
News/Talk	15.7
Oldies	7.1
Religious	14.6
	6.3
nguage	5.4
Other Form	11.4
Total	**99.9**

Percent

Displaying categorical data

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

Frequency Table

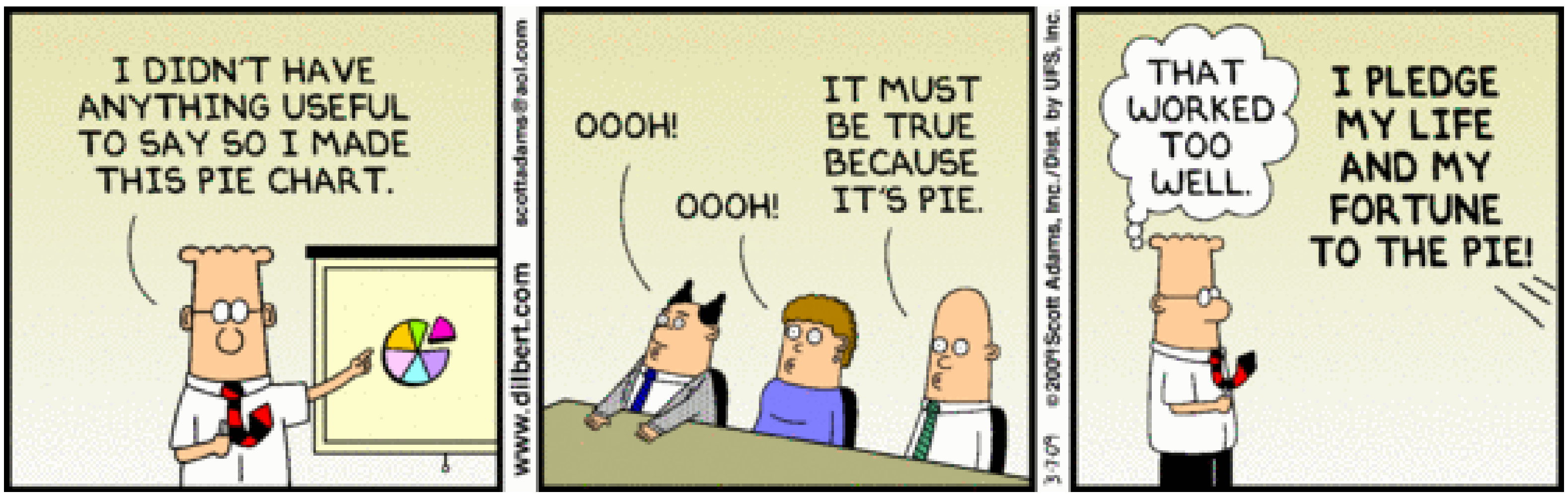
Format	Count of Stations
Adult Contemporary	1556
Adult Standards	1196
Contemporary Hit	569
Country	2066
News/Talk	2179
Oldies	1060
Religious	2014
Rock	869
Spanish Language	750
Other Formats	1579
Total	**13838**

Relative Frequency Table

Format	Percent of Stations
Adult Contemporary	11.2
Adult Standards	8.6
Contemporary Hit	4.1
Country	14.9
News/Talk	15.7
Oldies	7.7
Religious	14.6
Rock	6.3
Spanish Language	5.4
Other Formats	11.4
Total	**99.9**

Graphs: Good and Bad

Bar graphs compare several quantities by comparing the heights of bars that represent those quantities.



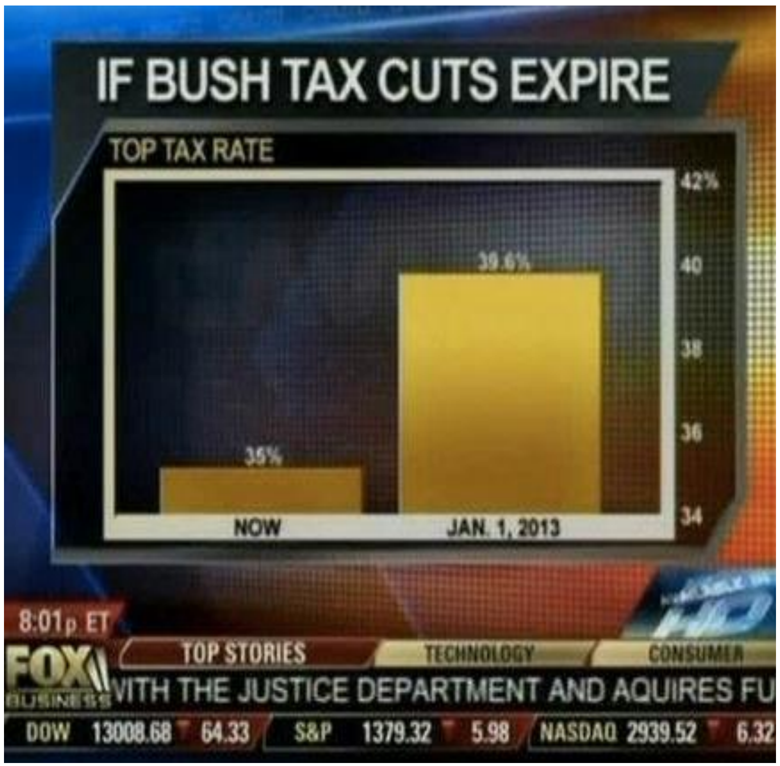
Alternate Example

This ad for DIRECTV has multiple problems. How many can you point out?

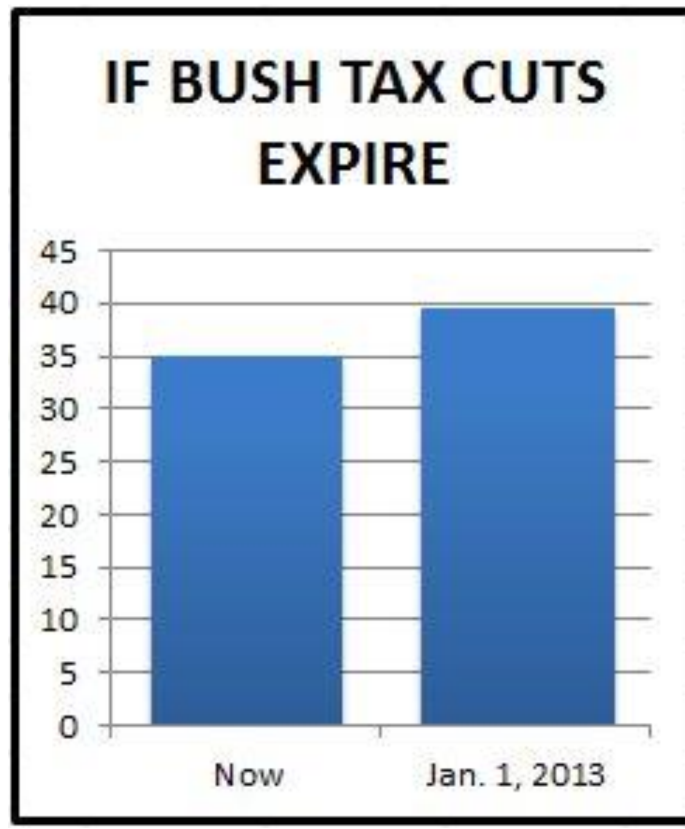
Alternate Example

This ad for DIRECTV has multiple problems. How many can you point out?

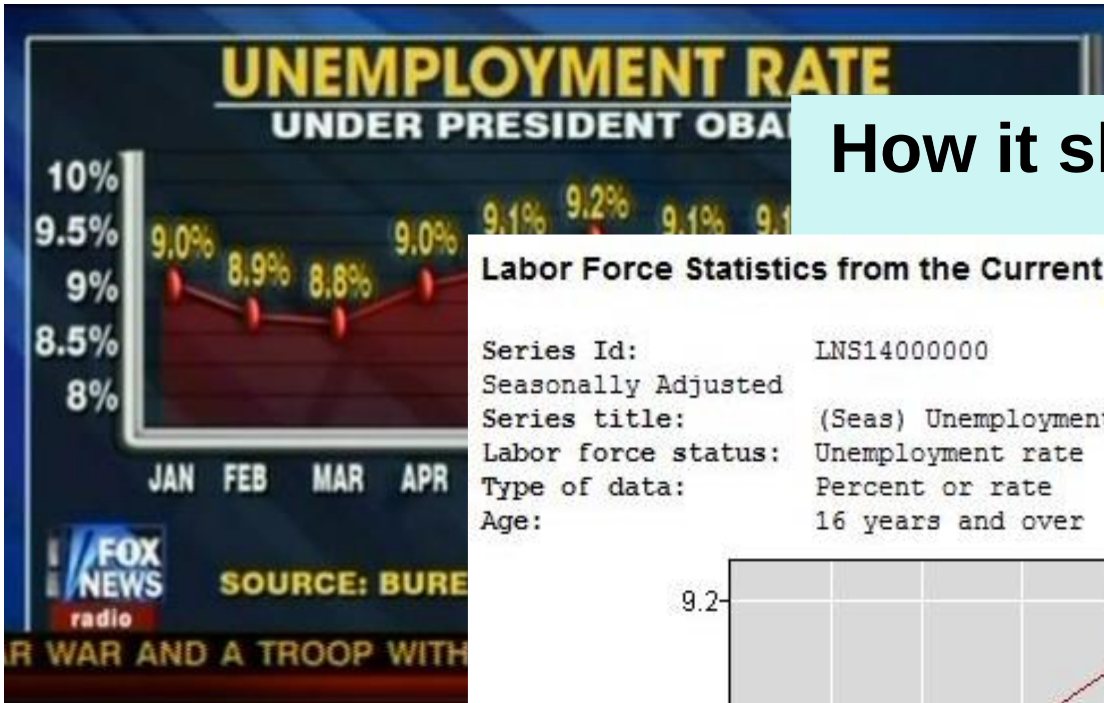
Examples of Misleading Statistics



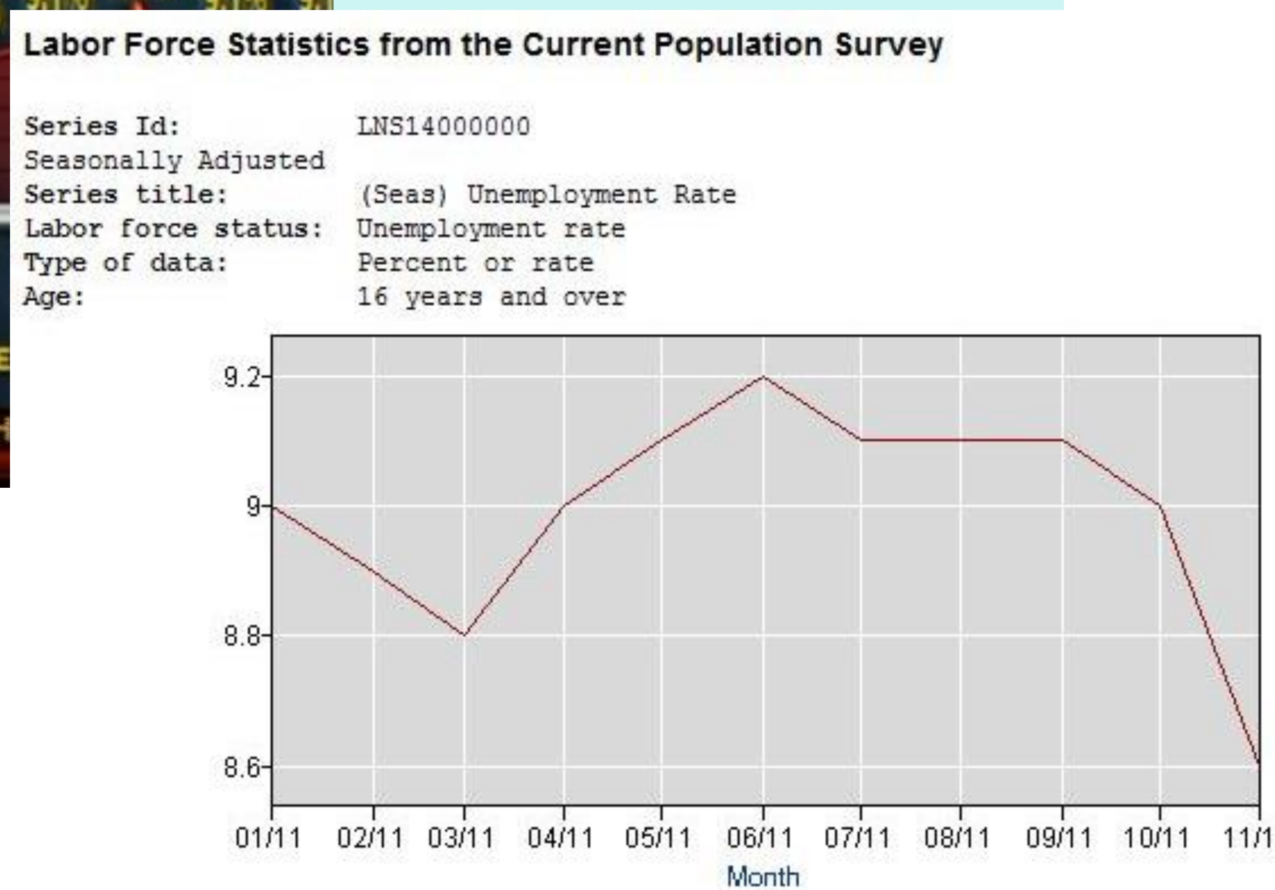
How it should look:



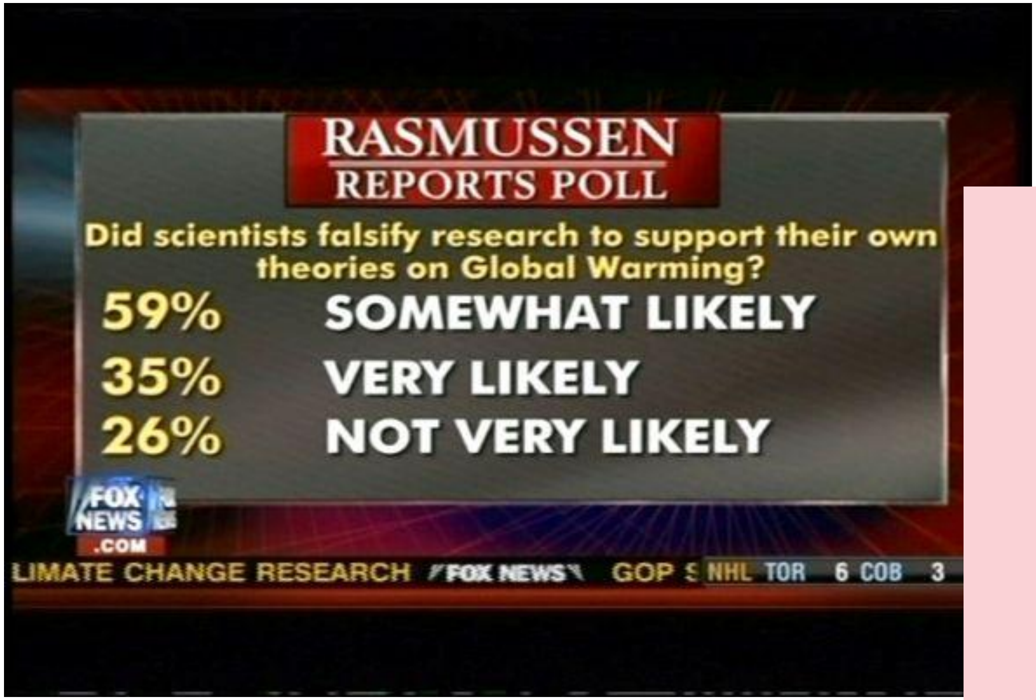
Examples of Misleading Statistics



How it should look:



Examples of Misleading Statistics



Problems??

- Can respondents vote twice?
- Rasmussen believes in giving more than 100%
- FOX fact checkers stink at math

Two-Way Tables and Marginal Distributions

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

Definition:

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

Example

Young adults by gender and chance of getting rich

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

What are the variables described by this two-way table?

Opinion & gender

How many young adults were surveyed?

4826 total

Two-Way Tables and Marginal Distributions

Definition:

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

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

To examine a marginal distribution,
1) Use the data in the table to calculate the marginal distribution (in *percents*) of the **row** or ***column*** totals.
2) Make a graph to display the marginal distribution.

Two-Way Tables and Marginal Distributions

Example

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

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

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

Relationships Between Categorical Variables

- Marginal distributions tell us nothing about the relationship between two variables.

Definition:

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

To examine or compare conditional distributions,

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

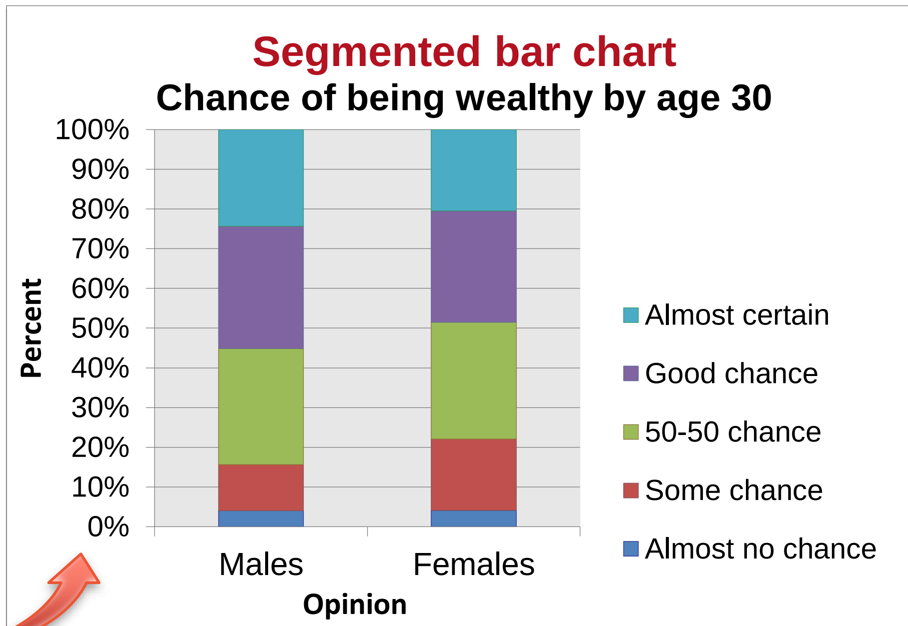
Two-Way Tables and Conditional Distributions

Example

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

Calculate the conditional distribution of opinion among males.
Examine the relationship between gender and opinion.

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



How would you complete this 2-way table for our class? (Last Year's R2)

Students by gender and Magnet in AP Stats			
	Female	**Male**	**TOTAL**
HSU	**2**	**2**	4
J & C	**3**	**2**	5
MST	**4**	**8**	12
VA	**2**	**0**	2
YPAS	**1**	**2**	3
TOTAL	**12**	**14**	**26**

How would you complete this 2-way table for our class? RED 4

Students by grade and Magnet in AP Stats				
	Soph.	Junior	Senior	TOTAL
HSU				
J & C				
MST				
VA				
YPAS				
TOTAL				

AP Stats Students 2020
(2-Way Frequency Table)

Joint Freq.

	Soph.	**Junior**	**Senior**	**TOTAL**
HSU	**2**	**2**	**9**	13
J & C	**0**	**1**	**2**	3
MST	**18**	**4**	**9**	31
VA	**0**	**1**	**1**	2
YPAS	**0**	**1**	**1**	2
TOTAL	**20**	**9**	**22**	**51**

Marginal Freq.

AP Stats Students 2020
(2-Way Relative Frequency Table)

	Soph.	Junior	Senior	TOTAL
HSU	**0.039**	**0.039**	**0.176**	0.254
J & C	**0**	**0.020**	**0.039**	0.059
MST	**0.353**	**0.078**	**0.176**	0.610
VA	**0**	**0.020**	**0.020**	0.039
YPAS	**0**	**0.020**	**0.020**	0.039
TOTAL	**0.392**	**0.176**	**0.431**	**1.000**

+

Section 3.1
Analyzing Categorical Data

Summary

In this section, we learned that…

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

Section 3.1 (& from Chapter 1) Analyzing Categorical Data

Summary, continued

In this section, we learned that…

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

Looking Ahead…

In the next Section…

We'll learn how to display quantitative data.

✓**Review Dotplots**

✓**Introduce Stemplots**

✓**Introduce Histograms**

We'll also learn how to describe and compare distributions of quantitative data.

Section 3.2
Displaying Quantitative Data with Graphs

Learning Objectives

After this section, you should be able to…

- ✓ CONSTRUCT and INTERPRET dotplots, stemplots, and histograms
- ✓ DESCRIBE the shape of a distribution
- ✓ COMPARE distributions
- ✓ USE histograms wisely

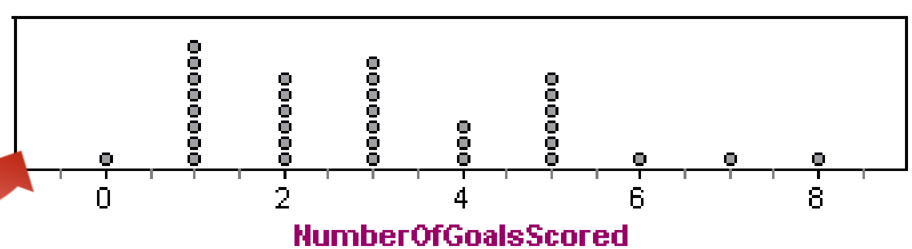
Dotplots

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

How to Make a Dotplot

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

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



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

How to Examine the Distribution of a Quantitative Variable

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

Describe the overall pattern of a distribution by its:

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

Don’t forget your SOCS!

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

Describing Shape

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

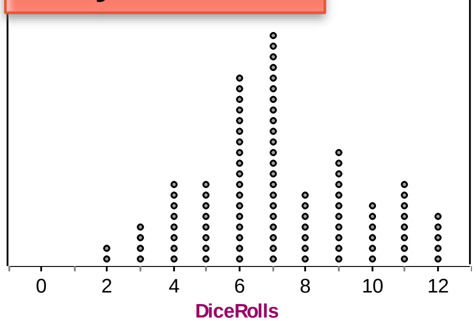
Definitions:

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

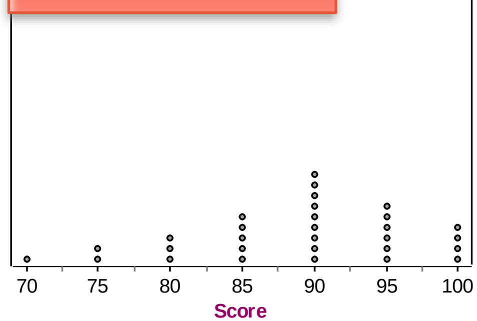
A distribution is **skewed to the right** (right-skewed or *positively skewed*) if the right side of the graph (containing the half of the observations with larger values) is much longer than the left side.

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

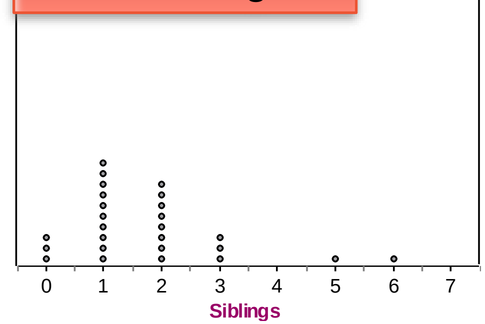
Symmetric



Skewed-left



Skewed-right



Distribution Shape

U.S. Income Distribution from 2005

Skewed Right or *positively skewed* →

Measuring Center: The Mean

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

Definition:

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

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

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

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

Measuring Center: The Median

- Another common measure of center is the **median**. In section 1.2, we learned that the median describes the midpoint of a distribution.

Definition:

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

To find the median of a distribution:

1) Arrange all observations from smallest to largest.

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

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

Find the mean of this mean & median of this data:

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

CENTERS (measures of central tendency): *mean, median, mode*

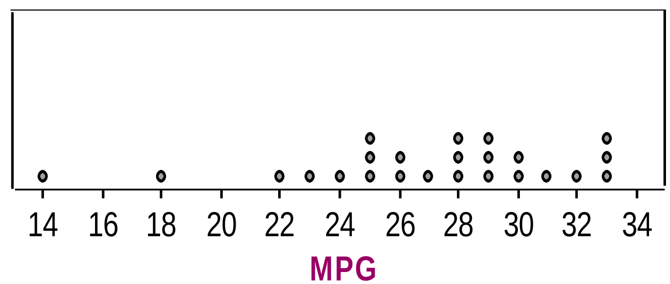
CENTER: *Mean = 26.958 mpg,*
Median = 28 mpg, Mode = multimodal

- **Examine this data**

Example

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

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



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

SHAPE: *skewed left or neg. skewed*

CENTER: *mean, Median, mode?*

SPREAD: *range is 19* (33 – 14)

Comparing the Mean and the Median

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

Comparing the Mean and the Median

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

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

In a skewed distribution, the mean is usually farther out in the long tail than is the median.

Stemplots (Stem-and-Leaf Plots)

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

How to Make a Stemplot

1) Separate each observation into a **stem** (all but the final digit) and a **leaf** (the final digit).

2) Write all possible stems from the smallest to the largest in a vertical column and draw a vertical line to the right of the column.

3) Write each leaf in the row to the right of its stem.

4) Arrange the leaves in increasing order out from the stem.

5) Provide a key that explains in context what the stems and leaves represent.

Measuring Center

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

Example

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

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

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

Key: 4|5 represents a New York worker who reported a 45-minute travel time to work.

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

Stemplots (Stem-and-Leaf Plots)

- 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	
2	
3	
4	
5	

Stems

1	93335
2	664233
3	1840
4	9
5	0701

Add leaves

1	33359
2	233466
3	0148
4	9
5	0017

Order leaves

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

Add a key

Splitting Stems and Back-to-Back Stemplots

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

Females

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

Males

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

Stem	
0	
0	
1	
1	
2	
2	"split stems"
3	
3	
4	
4	
5	
5	

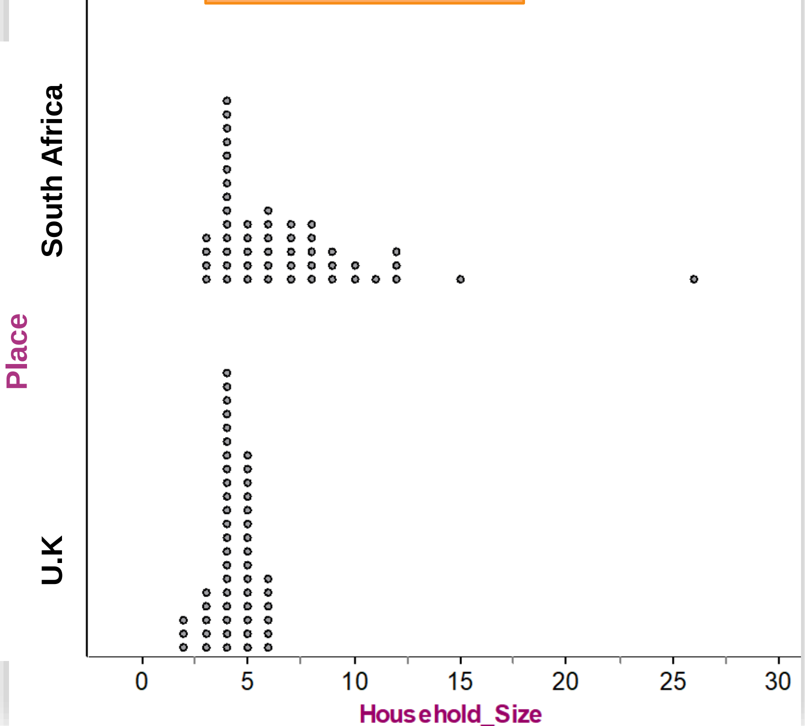
Females		Males
	0	4
	0	555677778
333	1	0000124
95	1	
4332	2	2
66	2	
410	3	
8	3	58
	4	
9	4	
100	5	
7	5	

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

Comparing Distributions

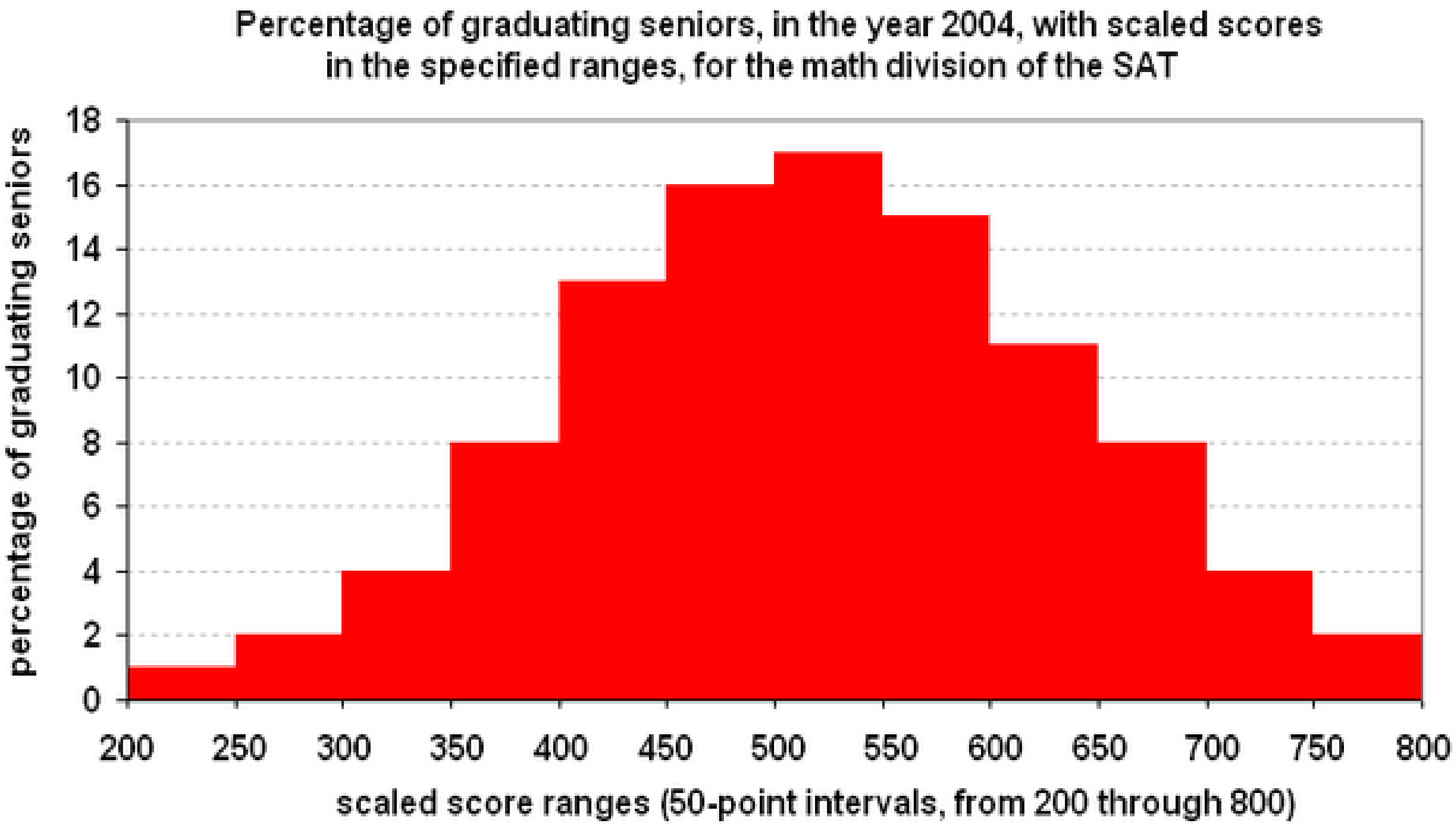
- Some of the most interesting statistics questions involve comparing two or more groups.
- Always discuss **shape**, **center**, **spread**, and possible **outliers** whenever you compare distributions of a quantitative variable.

Example



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

Chapter 3.3 - Numerical Data: Frequency Distributions using Histograms



Histograms

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

How to Make a Histogram

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

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

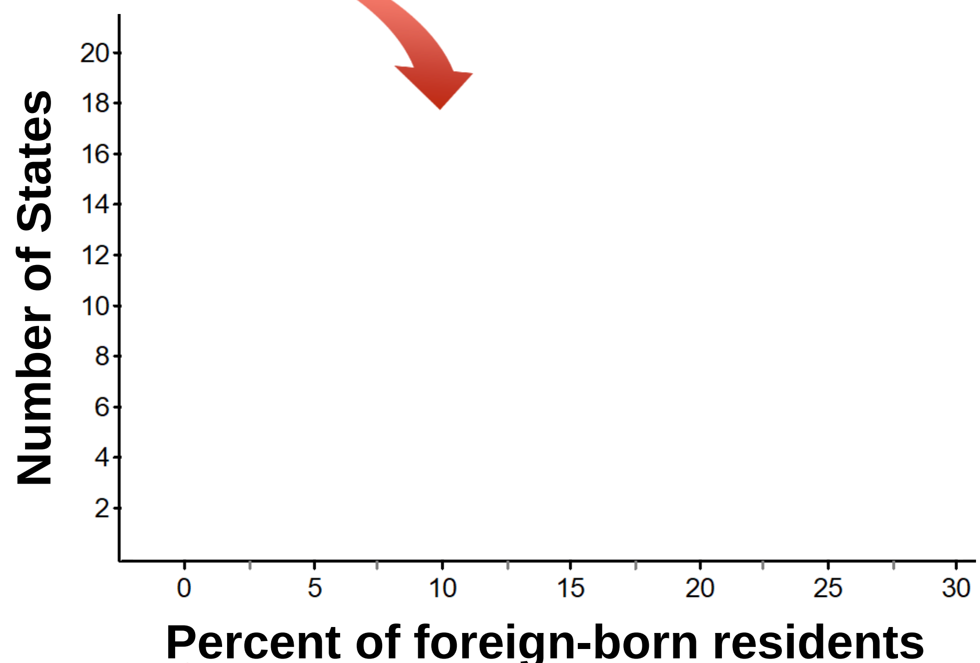
3) Label and scale your axes and draw the histogram. The height of the bar equals its frequency. Adjacent bars should touch, *unless* a class contains no individuals(observations).

Example

Making a Histogram

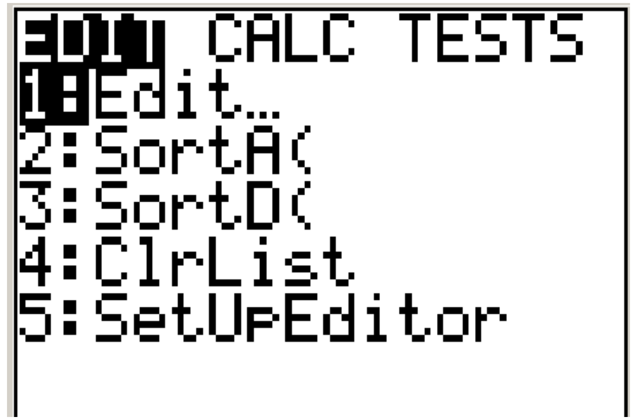
- The table below presents data on the percent of residents from each state who were born outside of the U.S.

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

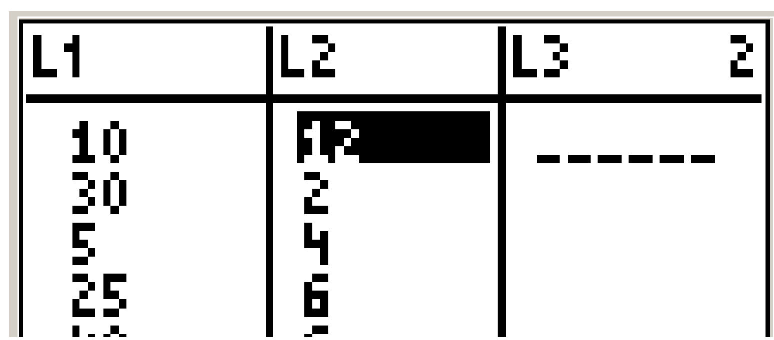


Practice Example: Promiscuous Queen Bees (p.99)

- Use your TI-84 calculator to input the data regarding Queen Bees and their number of partners during flight.

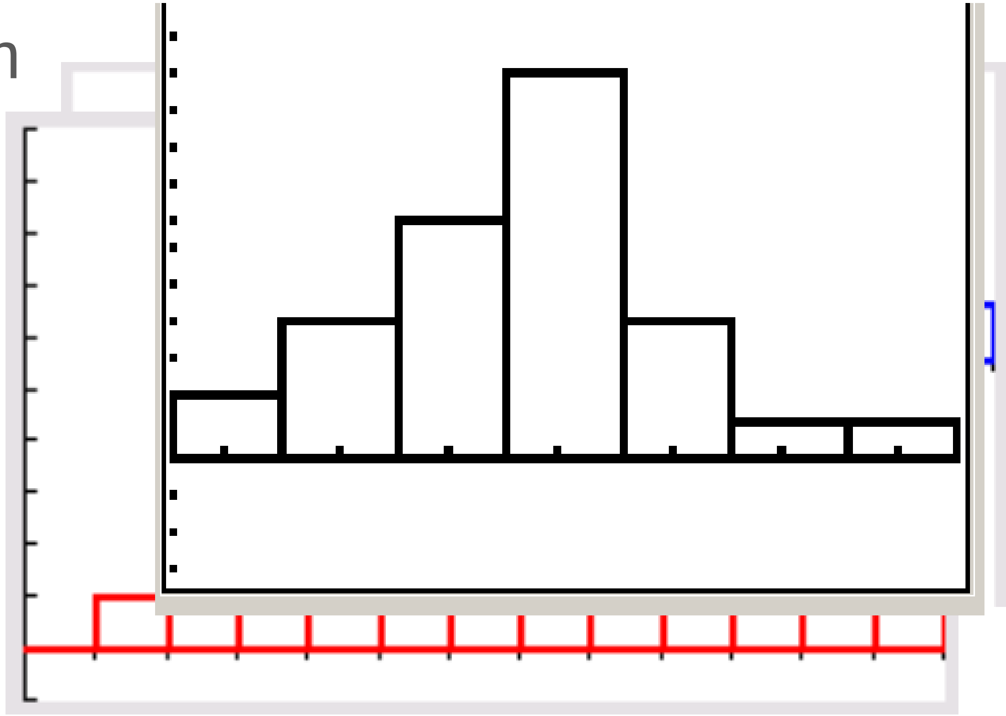


- From this data, we want to generate a histogram to graphically represent the data.



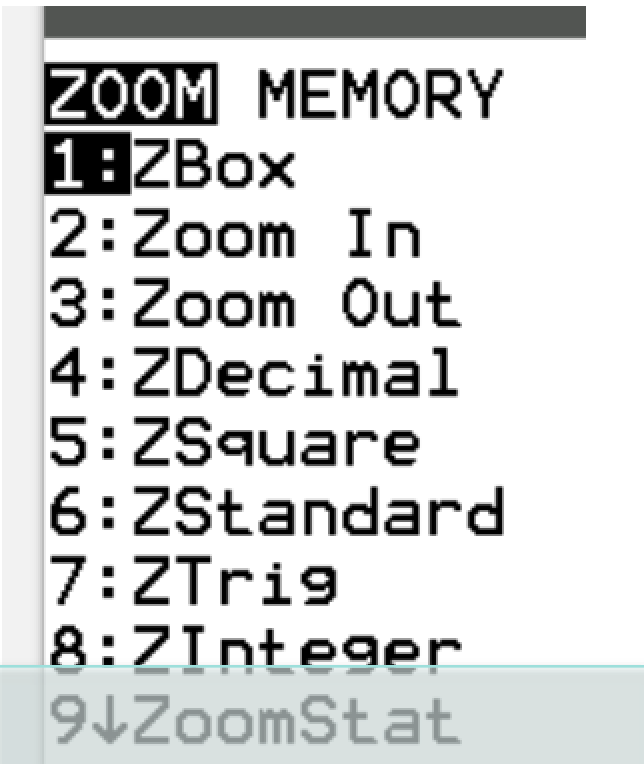
Practice Example: Promiscuous Queen Bees (p.99)

- Generate a Histogram with you TI-84 calculators
- What is your window setting?
- Can you change the intervals?
- Questions

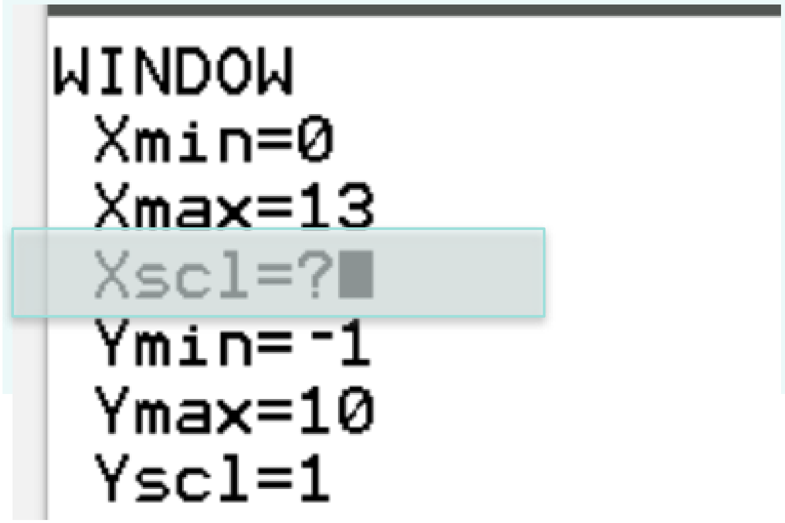


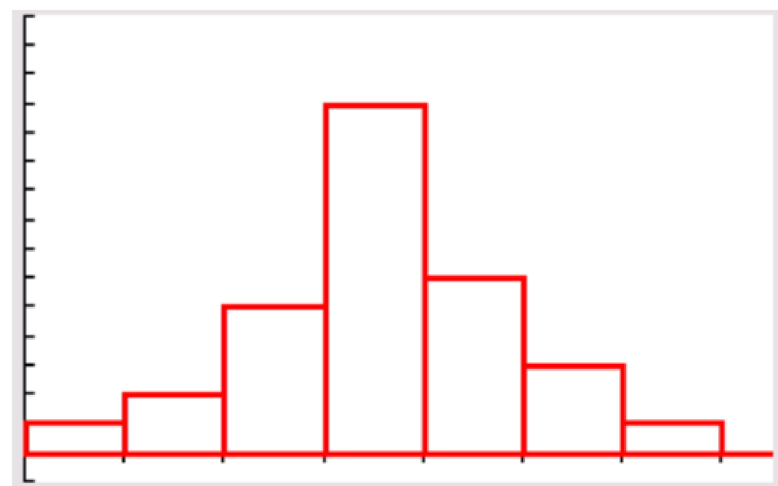
Histograms on TI-84

- You can change the **zoom**:



- You can change the **intervals**:





Bar Chart or Histogram?

Prevention and Management of Calcaneal Apophysitis Children: An Overuse Syndrome.
Lyle J. Micheli, M.D., and M. Lloyd Ireland, M.D.
Journal of Pediatric Orthopedics 7:34 - 38 © 1987 Raven Press, New York

Using Histograms Wisely

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

Cautions

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

2) When comparing distributions with different numbers of observations (different size samples) use *percents* instead of counts on the vertical axis (*relative frequency*).

3) Choose the best graphical display**:** *bar chart, dot plot, or histogram***.** Just because a graph looks nice, it’s not necessarily a meaningful display of data.