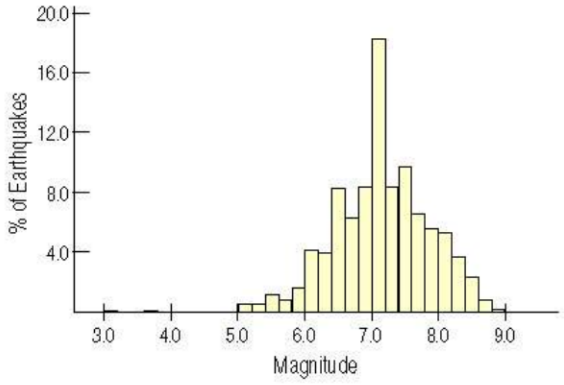
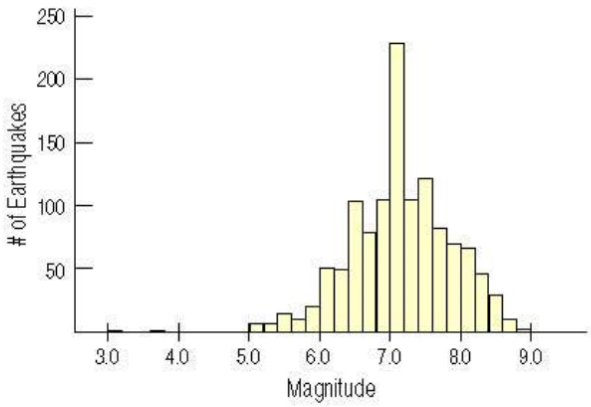
Histograms: Displaying the Distribution of Earthquake Magnitudes (cont.)

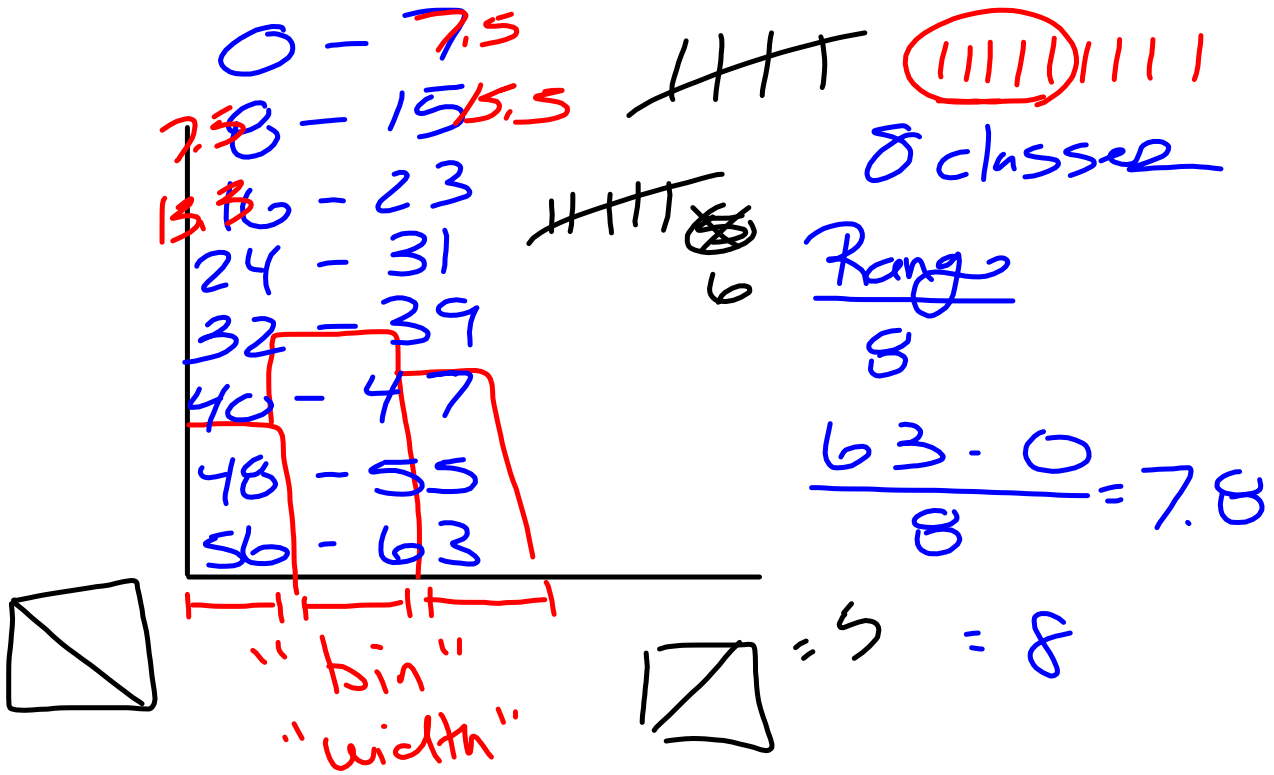
- A relative frequency histogram displays the *percentage* of cases in each bin instead of the count.
  - In this way, relative frequency histograms are faithful to the area principle.
- Here is a relative frequency histogram of earthquake magnitudes:



Histograms: Displaying the Distribution of Earthquake Magnitudes (cont.)

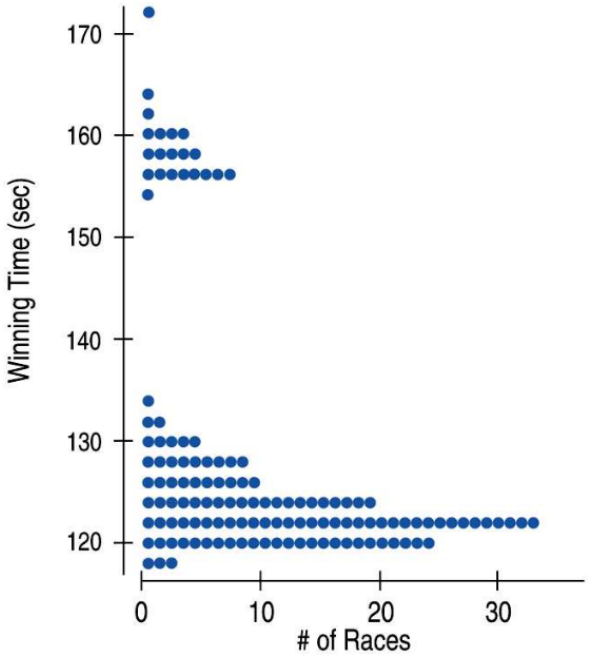
- A histogram plots the bin counts as the heights of bars (like a bar chart).
- It displays the distribution at a glance.
- Here is a histogram of earthquake magnitudes:




0 - 7.5
7.5 - 15.5
15.5 - 23
24 - 31
32 - 39
40 - 47
48 - 55
56 - 63
"bin"
"width"
HHT
HHT
6
8 classes
Range
63.0
8
= 7.8
: 5 = 8

Dotplots

- A dotplot is a simple display. It just places a dot along an axis for each case in the data.
- The dotplot to the right shows Kentucky Derby winning times, plotting each race as its own dot.
- You might see a dotplot displayed horizontally or vertically.



Pg. 44

What is a distribution?

Pg. 45

What is the difference between a frequency table and a relative frequency table? When is it better to use relative frequency tables?

Pg. 44-45

Histograms (half-day)

The following table presents the average points scored per game (PPG) for the 30 NBA teams in the 2009–2010 regular season. Make a dotplot to display the distribution of points per game. Then, use your dotplot to make a histogram of the distribution.

Team	PPG	Team	PPG	Team	PPG
Atlanta Hawks	101.7	Indiana Pacers	100.8	Oklahoma City Thunder	101.5
Boston Celtics	99.2	Los Angeles Clippers	95.7	Orlando Magic	102.8
Charlotte Bobcats	95.3	Los Angeles Lakers	101.7	Philadelphia 76ers	97.7
Chicago Bulls	97.5	Memphis Grizzlies	102.5	Phoenix Suns	110.2
Cleveland Cavaliers	102.1	Miami Heat	96.5	Portland Trail Blazers	98.1
Dallas Mavericks	102	Milwaukee Bucks	97.7	Sacramento Kings	100
Denver Nuggets	106.5	Minnesota Timberwolves	98.2	San Antonio Spurs	101.4
Detroit Pistons	94	New Jersey Nets	92.4	Toronto Raptors	104.1
Golden State Warriors	108.8	New Orleans Hornets	100.2	Utah Jazz	104.2
Houston Rockets	102.4	New York Knicks	102.1	Washington Wizards	96.2

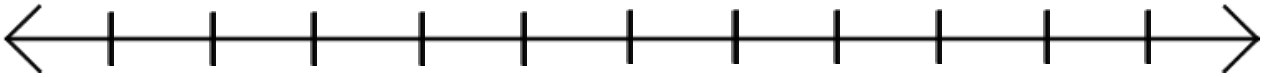
How do you make a histogram?

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

Histograms (half-day)

The following table presents the average points scored per game (PPG) for the 30 NBA teams in the 2009–2010 regular season. Make a dotplot to display the distribution of points per game. Then, use your dotplot to make a histogram of the distribution.

Team	PPG	Team	PPG	Team	PPG
Atlanta Hawks	101.7	Indiana Pacers	100.8	Oklahoma City Thunder	101.5
Boston Celtics	99.2	Los Angeles Clippers	95.7	Orlando Magic	102.8
Charlotte Bobcats	95.3	Los Angeles Lakers	101.7	Philadelphia 76ers	97.7
Chicago Bulls	97.5	Memphis Grizzlies	102.5	Phoenix Suns	110.2
Cleveland Cavaliers	102.1	Miami Heat	96.5	Portland Trail Blazers	98.1
Dallas Mavericks	102	Milwaukee Bucks	97.7	Sacramento Kings	100
Denver Nuggets	106.5	Minnesota Timberwolves	98.2	San Antonio Spurs	101.4
Detroit Pistons	94	New Jersey Nets	92.4	Toronto Raptors	104.1
Golden State Warriors	108.8	New Orleans Hornets	100.2	Utah Jazz	104.2
Houston Rockets	102.4	New York Knicks	102.1	Washington Wizards	96.2



EXAMPLE Goooooaaaaallllll!

How to make a dotplot

How good was the 2004 U.S. women's soccer team? With players like Brandi Chastain, Mia Hamm, and Briana Scurry, the team put on an impressive showing en route to winning the gold medal at the 2004 Olympics in Athens. Here are data on the number of goals scored by the team in 34 games played during the 2004 season:[20]

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

Here are the steps in making a dotplot:

- *Draw a horizontal axis (a number line) and label it with the variable name.* In this case, the variable is number of goals scored.
- *Scale the axis.* Start by looking at the minimum and maximum values of the variable. For these data, the minimum number of goals scored was 0, and the maximum was 8. So we mark our scale from 0 to 8, with tick marks at every whole-number value.
- *Mark a dot above the location on the horizontal axis corresponding to each data value.* Figure 1.9 displays a completed dotplot for the soccer data.

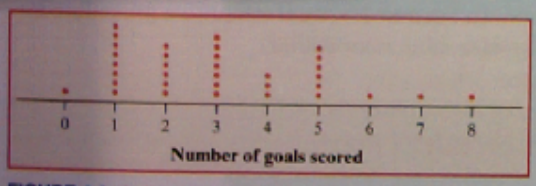
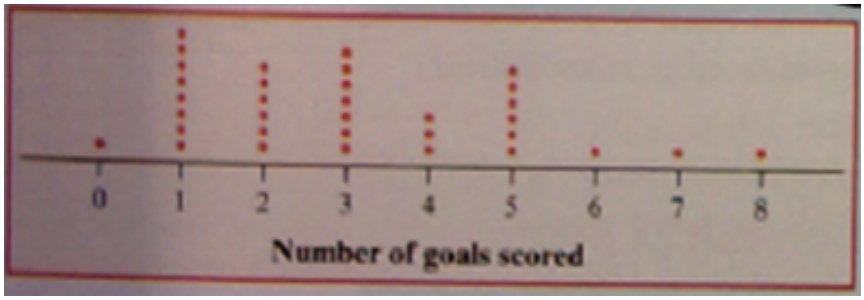


FIGURE 1.9 A dotplot of goals scored by the U.S. women's soccer team in 2004.



Shape: The dotplot has a peak at 1. This indicates that the team's most frequent number of goals scored in games that season (known as the **mode**) was 1. In most of its games, the U.S. women's soccer team scored between 1 and 5 goals. However, the distribution has a long tail to the right. (Later, we will describe the shape of Figure 1.9 as *skewed to the right.*)

Center: We can describe the center by finding a value that divides the observations so that about half take larger values and about half take smaller values. This value is called the *median* of the distribution. In Figure 1.9, the median is 3. That is, in a typical game during the 2004 season, the U.S. women's soccer team scored about 3 goals. Of course, we could also summarize the center of the distribution by calculating the average (*mean*) number of goals scored per game. For the 2004 season, the team's mean was 3.06 goals.

Spread: The spread of a distribution tells us how much *variability* there is in the data. One way to describe the variability is to give the smallest and largest values. The spread in Figure 1.9 is from 0 goals to 8 goals scored. Alternatively, we can compute the **range** of the distribution by subtracting the smallest value from the largest value. For these data, the range is $8 - 0 = 8$ goals.

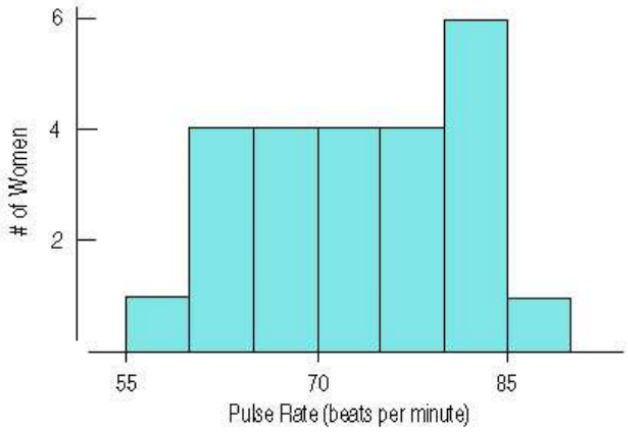
Outliers: Was the game in which the women's team scored 8 goals an outlier? How about the team's 7-goal game? These values differ somewhat from the overall pattern. However, they don't clearly stand apart from the rest of the distribution. For now, let's agree to call attention only to potential outliers that suggest something special about an observation. In Section 1.3, we'll establish a procedure for determining whether a particular data value is an outlier.

Stem-and-Leaf Displays

- Stem-and-leaf displays show the distribution of a quantitative variable, like histograms do, while preserving the individual values.
- Stem-and-leaf displays contain all the information found in a histogram and, when carefully drawn, satisfy the area principle and show the distribution.

Stem-and-Leaf Example

- Compare the histogram and stem-and-leaf display for the pulse rates of 24 women at a health clinic. Which graphical display do *you* prefer?



8	8
8	000044
7	6666
7	2222
6	8888
6	0444
5	6

Constructing a Stem-and-Leaf Display

- First, cut each data value into leading digits ("stems") and trailing digits ("leaves").
- Use the stems to label the bins.
- Use only one digit for each leaf—either round or truncate the data values to one decimal place after the stem.

Think Before You Draw, Again

- Remember the "Make a picture" rule?
- Now that we have options for data displays, you need to *Think* carefully about which type of display to make.
- Before making a stem-and-leaf display, a histogram, or a dotplot, check the
  - Quantitative Data Condition: The data are values of a quantitative variable whose units are known.

Shape, Center, and Spread

- When describing a distribution, make sure to always tell about three things: shape, center, and spread…

What is the Shape of the Distribution?

1. Does the histogram have a single, central hump or several separated humps?
2. Is the histogram symmetric?
3. Do any unusual features stick out?

Pg. 47-48

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

Alternate Example: Which gender is taller, males or females? A sample of 14-year-olds from the United Kingdom was randomly selected using the CensusAtSchool website. Here are the heights of the students (in cm). Make a back-to-back stemplot and compare the distributions.

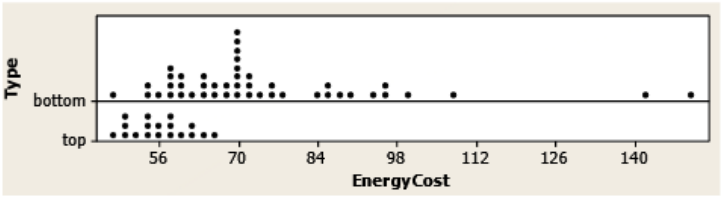
Male: 154, 157, 187, 163, 167, 159, 169, 162, 176, 177, 151, 175, 174, 165, 165, 183, 180

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

Pg. 49

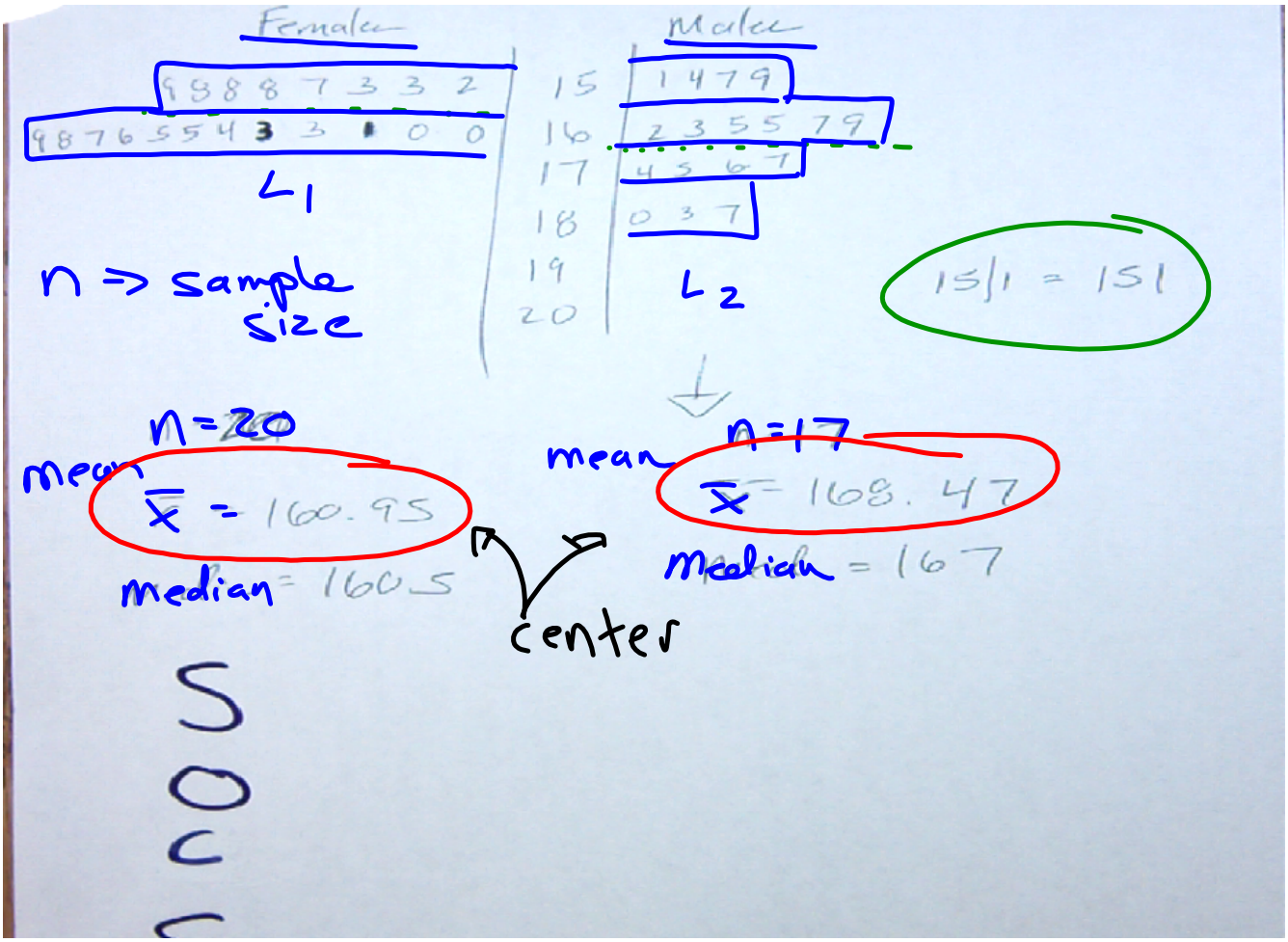
What is the most important thing to remember when you are asked to compare two distributions?

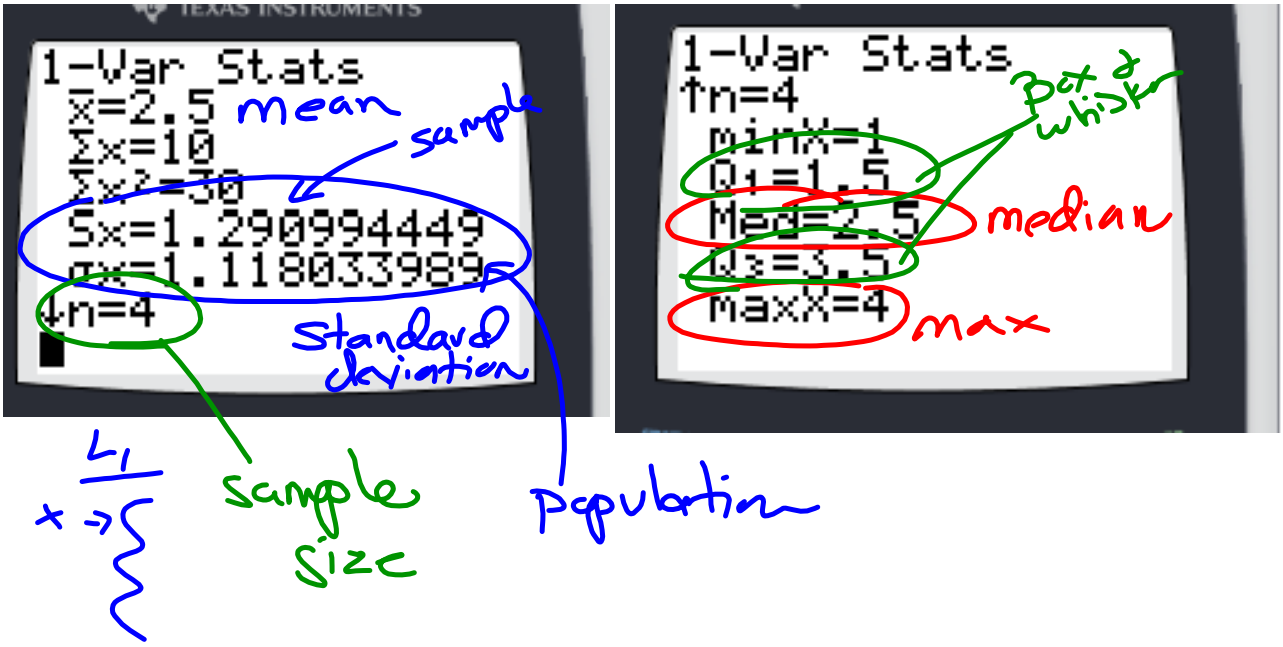
Alternate Example: Energy Cost: Top vs. Bottom Freezers
How do the annual energy costs (in dollars) compare for refrigerators with top freezers and refrigerators with bottom freezers? The data below is from the May 2010 issue of *Consumer Reports*.

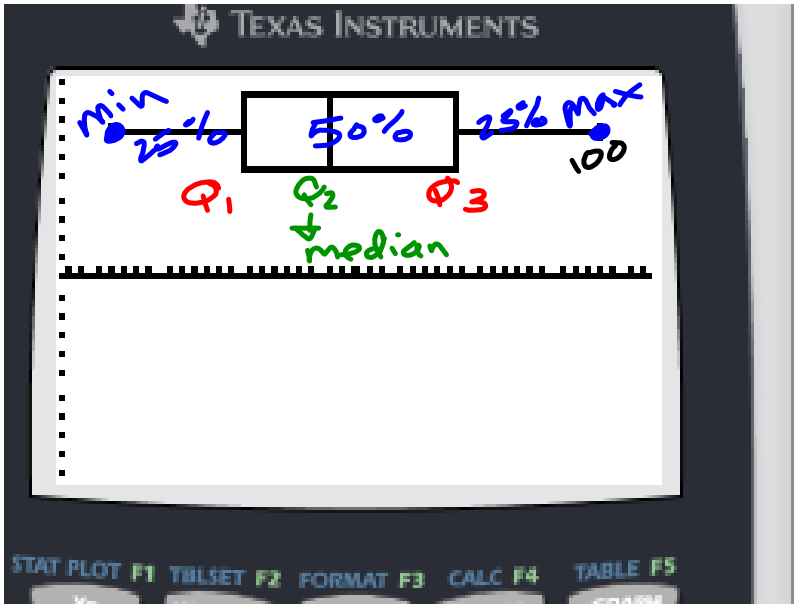


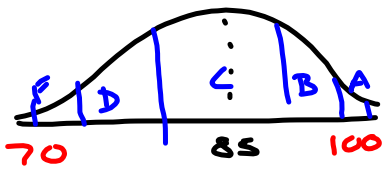
SOCS

S - Shape
O - Outliers
C - Center
S - Spread


Females
9 9 8 8 7 3 3 2
9 8 7 6 5 5 4 3 3 1 0 0
L1
n => sample size
Males
15 | 1 4 7 9
16 | 2 3 5 5 7 9
17 | 4 5 6 7
18 | 0 3 7
19
20
L2
15|1 = 151
n=20
mean
x̄ = 160.95
median = 160.5
n=17
mean
x̄ = 168.47
median = 167
center
S
O
C


1-Var Stats
x̄=2.5
Σx=10
Σx²=30
Sx=1.290994449
σx=1.118033989
n=4
mean
sample
standard deviation
1-Var Stats
n=4
minX=1
Q1=1.5
Med=2.5
Q3=3.5
maxX=4
box & whisker
median
max
L1
sample size
population


TEXAS INSTRUMENTS
min
25%
50%
25%
max
100
Q1
Q2
Q3
median
STAT PLOT F1
TBLSET F2
FORMAT F3
CALC F4
TABLE F5


F
D
C
B
A
70
85
100

2005 AP question #1a

The goal of a nutritional study was to compare the caloric intake of adolescents living in rural areas of the US with the caloric intake of adolescents living in urban areas of the United States. A random sample of ninth-grade students from one high school in a rural area was selected. Another random sample of ninth graders from one high school in an urban area was also selected. Each student in each sample kept records of all the food he or she consumed in one day.

The back-to-back stemplot below displays the number of calories of food consumed per kilogram of body weight for each student on that day.

Urban		Rural
99998876	2	
44310	3	2334
97665	3	56667
20	4	02224
	4	56889
	5	1

Stem: tens
Leaf: ones

2
3
4
5

(a) Write a few sentences comparing the distribution of the daily caloric intake of ninth-grade students in the rural high school with the distribution of daily caloric intake of ninth-grade students in the urban high school.

must be in context

SOCS

Collecting Data

Displaying Quantitative Data with Graphs

Pg. 49 – 50

When describing the distribution of a quantitative variable, what characteristics should be addressed?

Briefly describe/illustrate the following distribution shapes:

Symmetric | Skewed right | Skewed left

Unimodal | Bimodal | Uniform

Alternate Example: Smart Phone Battery Life
Here is the estimated battery life for each of 9 different smart phones (in minutes). Make a dotplot of the data and describe what you see.

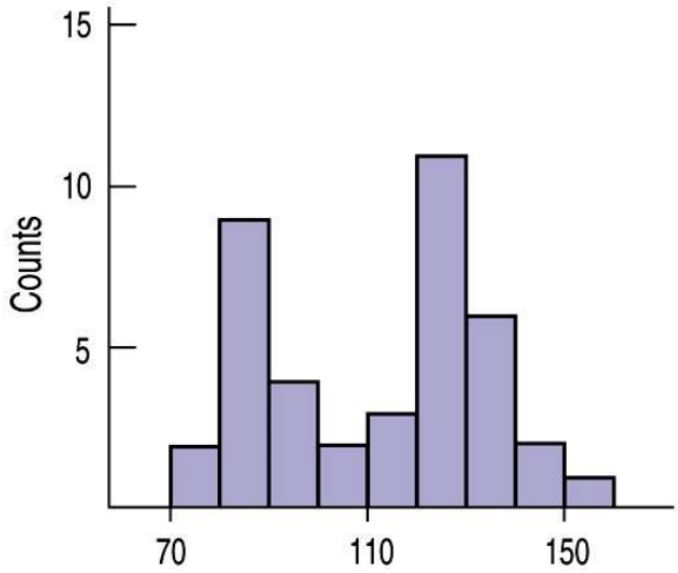
Smart Phone	Battery Life (minutes)
Apple iPhone	300
Motorola Droid	385
Palm Pre	300
Blackberry Bold	360
Blackberry Storm	330
Motorola Cliq	360
Samsung Moment	330
Blackberry Tour	300
HTC Droid	460

Humps

1. Does the histogram have a single, central hump or several separated bumps?
   - Humps in a histogram are called modes.
   - A histogram with one main peak is dubbed unimodal; histograms with two peaks are bimodal; histograms with three or more peaks are called multimodal.

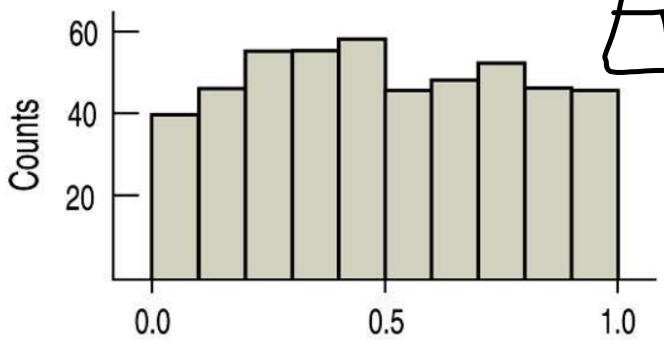
Humps (cont.)

- A bimodal histogram has two apparent peaks:



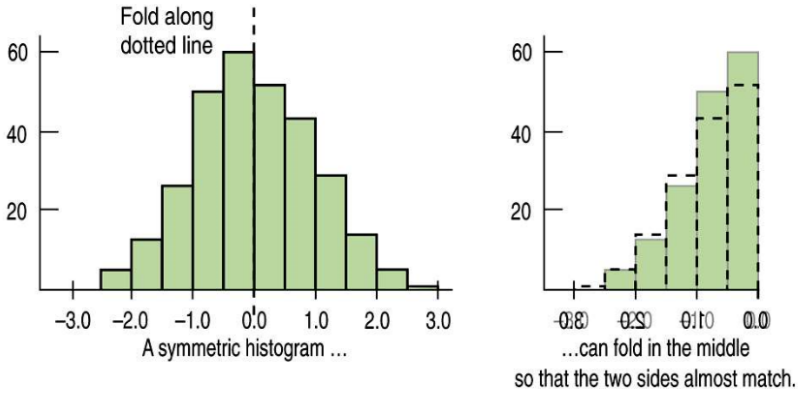
Humps (cont.)

- A histogram that doesn't appear to have any mode and in which all the bars are approximately the same height is called uniform:



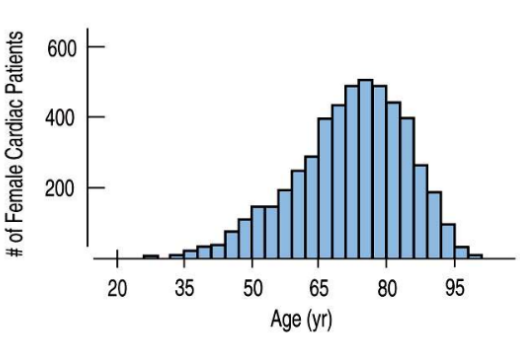
Symmetry

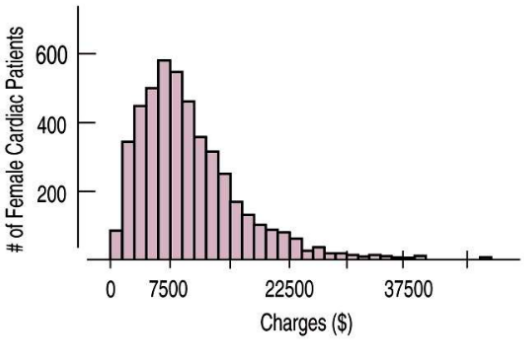
2. Is the histogram symmetric?
   - If you can fold the histogram along a vertical line through the middle and have the edges match pretty closely, the histogram is symmetric.



Symmetry (cont.)

- The (usually) thinner ends of a distribution are called the tails. If one tail stretches out farther than the other, the histogram is said to be skewed to the side of the longer tail.
- In the figure below, the histogram on the left is said to be skewed left, while the histogram on the right is said to be skewed right.



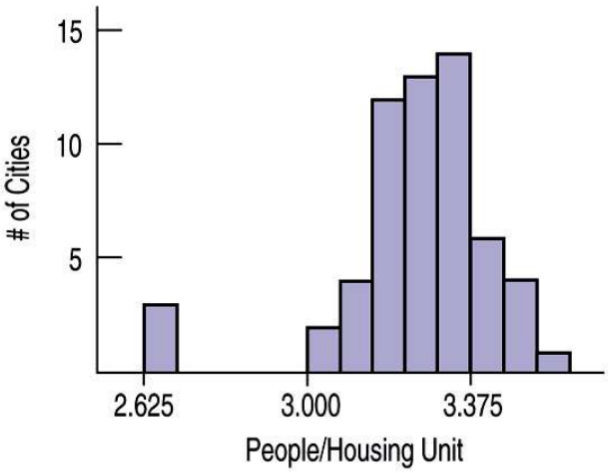


Anything Unusual?

3. Do any unusual features stick out?
   - Sometimes it's the unusual features that tell us something interesting or exciting about the data.
   - You should always mention any stragglers, or outliers, that stand off away from the body of the distribution.
   - Are there any gaps in the distribution? If so, we might have data from more than one group.

Anything Unusual? (cont.)

- The following histogram has outliers—there are three cities in the leftmost bar:

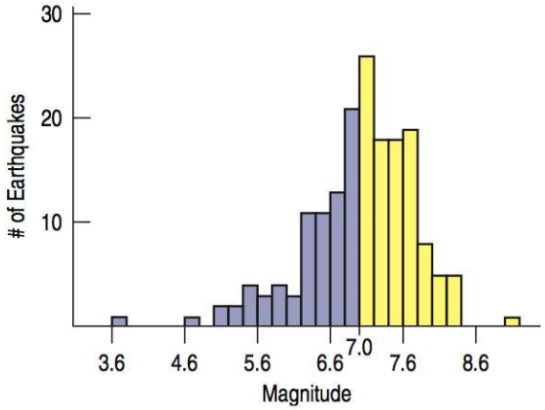


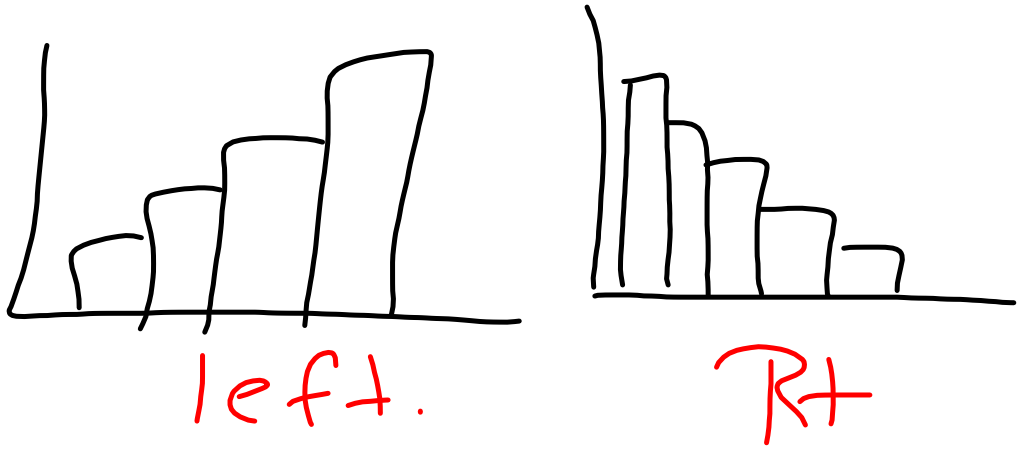
Where is the Center of the Distribution?

- If you had to pick a single number to describe all the data what would you pick?
- It's easy to find the center when a histogram is unimodal and symmetric—it's right in the middle.
- On the other hand, it's not so easy to find the center of a skewed histogram or a histogram with more than one mode.

Center of a Distribution -- Median

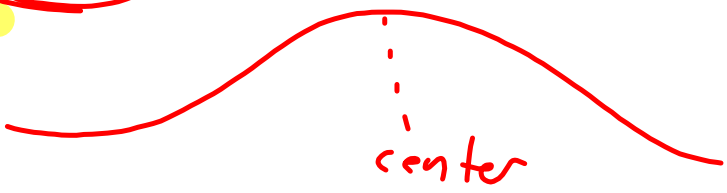
- The median is the value with exactly half the data values below it and half above it.
  - It is the middle data value (once the data values have been ordered) that divides the histogram into two equal areas
  - It has the same units as the data




left.
Rt

How Spread Out is the Distribution?

- Variation matters, and Statistics is about variation.
- Are the values of the distribution tightly clustered around the center or more spread out?
- Always report a measure of spread along with a measure of center when describing a distribution numerically.



Describing Quantitative Data with Numbers

Pg. 52 - 53

Is the median a resistant measure of center? Explain.

median when outliers

mean when no outliers

How does the shape of a distribution affect the relationship between the mean and the median?

bell shaped.

skewed

Pg. 54

What is the range? Is it a resistant measure of spread? Explain.

SOCS

spread

largest data value − smallest data value

What are quartiles? How do you find them?

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

$Q_3 - Q_1$ 50%

Pg. 59

What is the difference between $\bar{x}$ and μ?

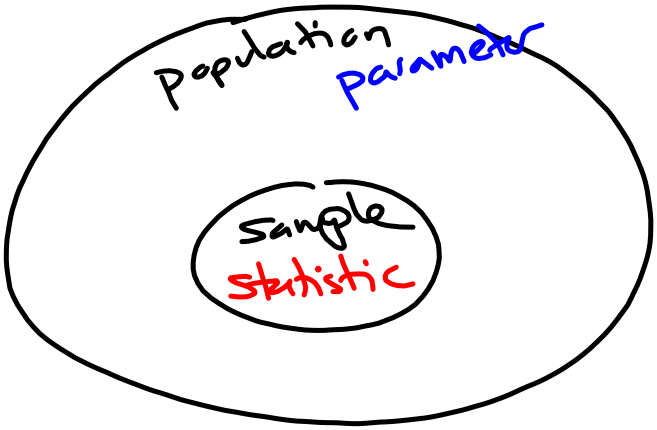
sample $\bar{x}$

population μ

} means

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

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


Population
parameter
sample
statistic