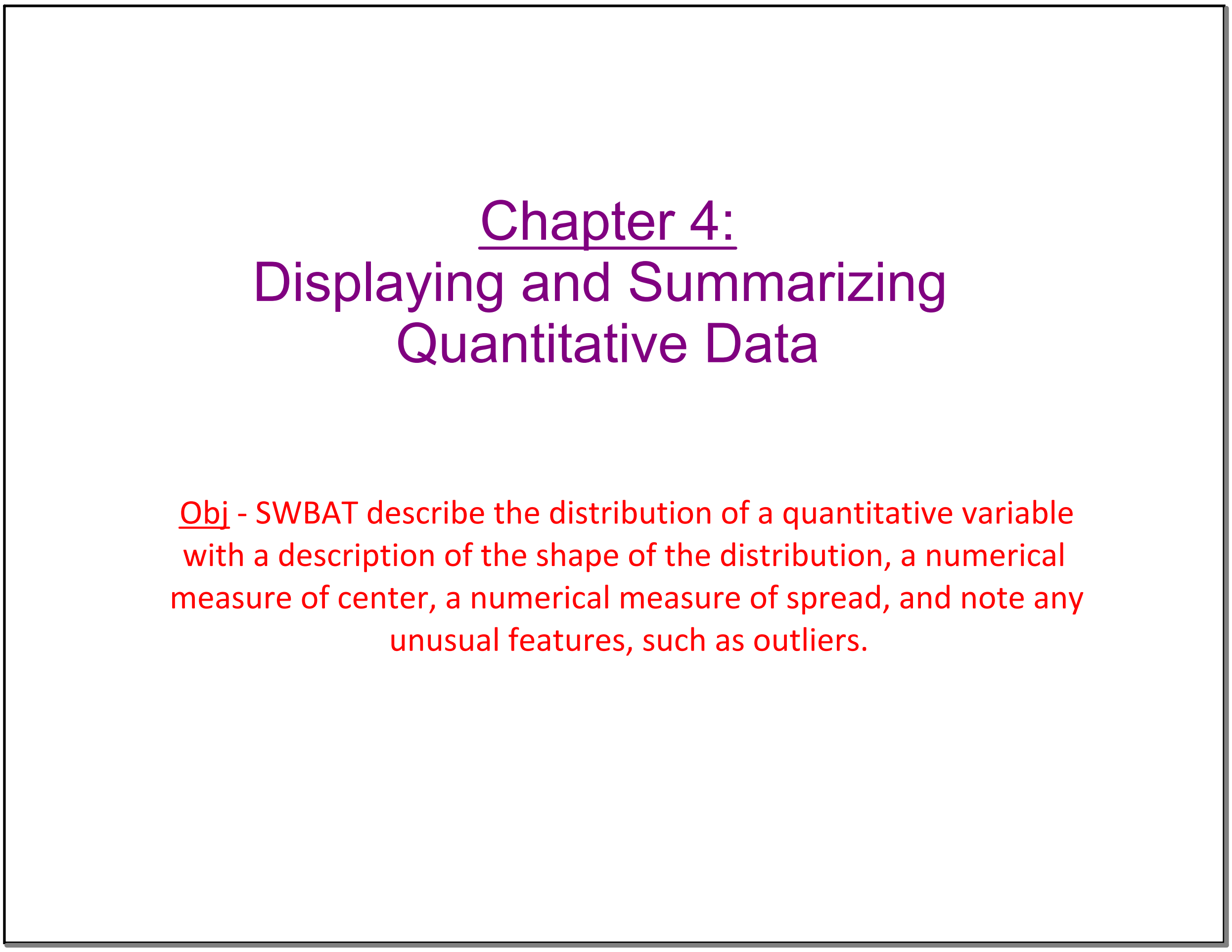
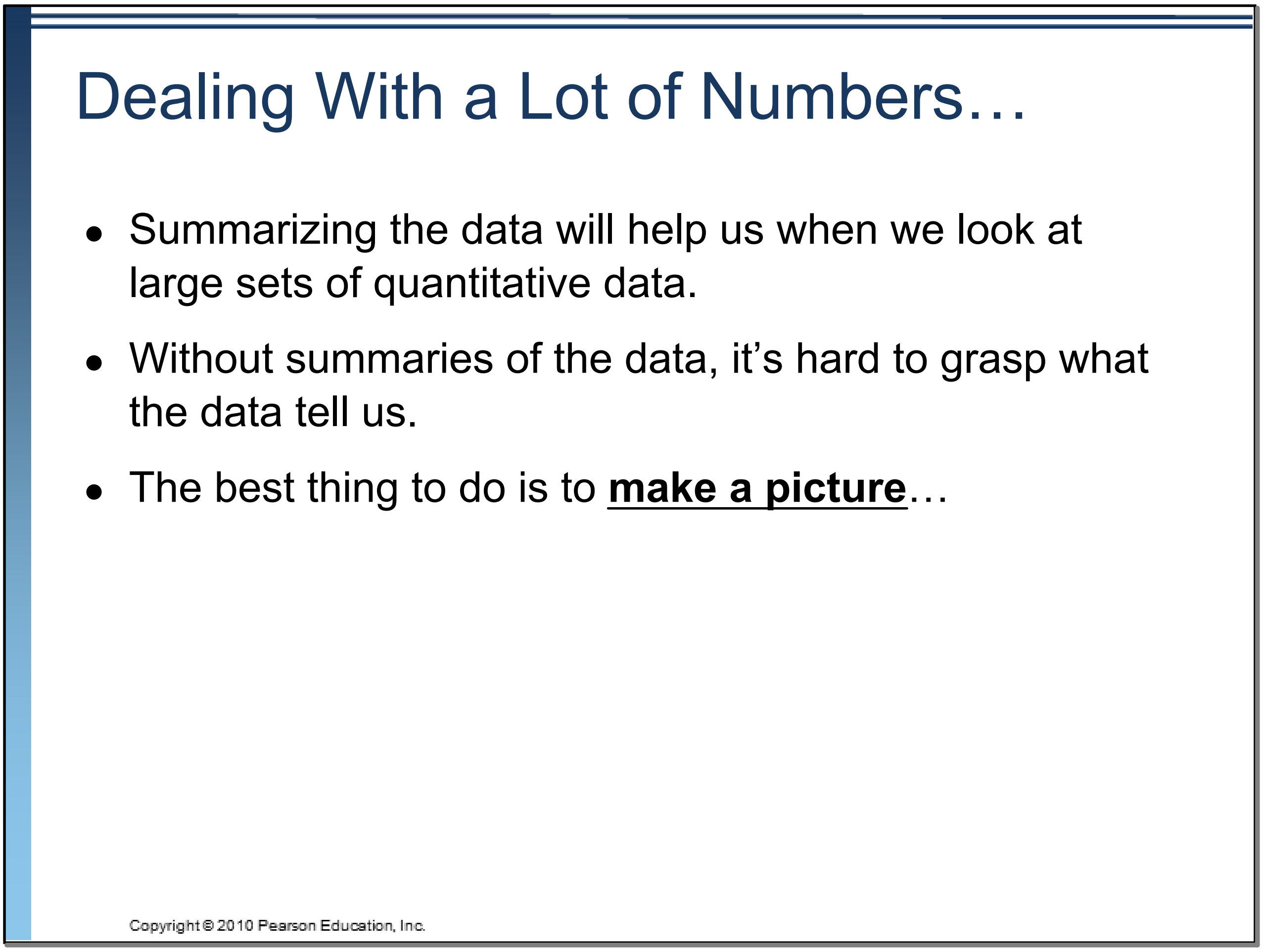


Chapter 4: Displaying and Summarizing Quantitative Data

Obj - SWBAT describe the distribution of a quantitative variable with a description of the shape of the distribution, a numerical measure of center, a numerical measure of spread, and note any unusual features, such as outliers.



Dealing With a Lot of Numbers…

- Summarizing the data will help us when we look at large sets of quantitative data.
- Without summaries of the data, it's hard to grasp what the data tell us.
- The best thing to do is to **make a picture**…

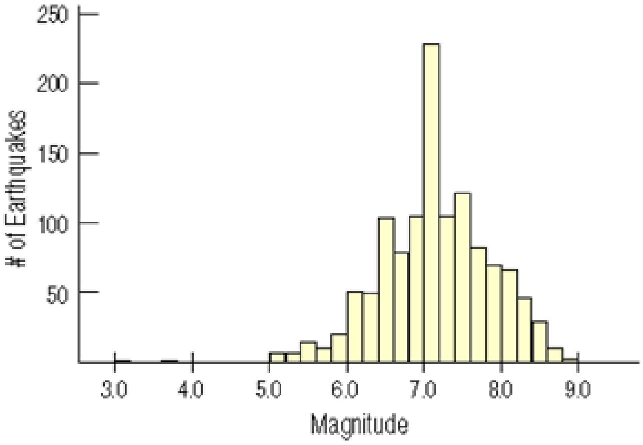
Histograms: Displaying the Distribution of Earthquake Magnitudes

The chapter example discusses earthquake magnitudes.

First, slice up the entire span of values covered by the quantitative variable into equal-width piles called bins.

The bins and the counts in each bin give the distribution of the quantitative variable.

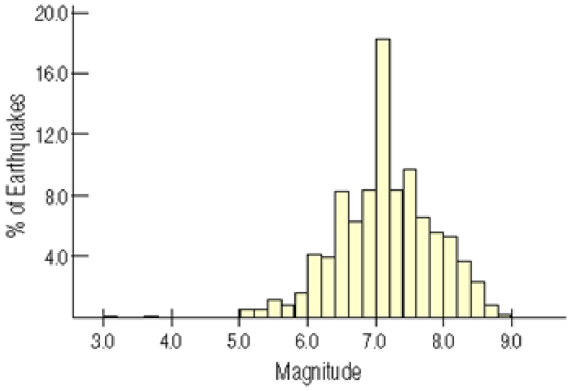
Histograms: Displaying the Distribution of Earthquake Magnitudes (cont.)



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

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:

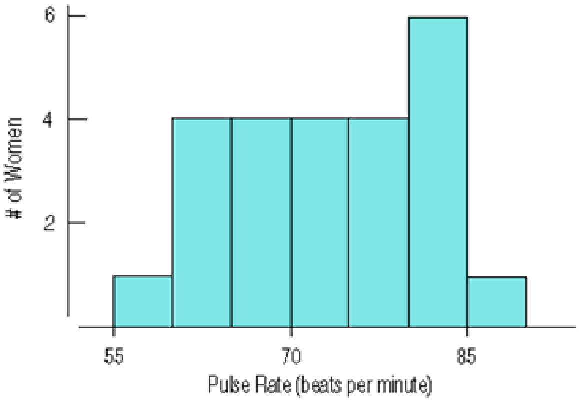


Stem-and-Leaf Displays

- 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?



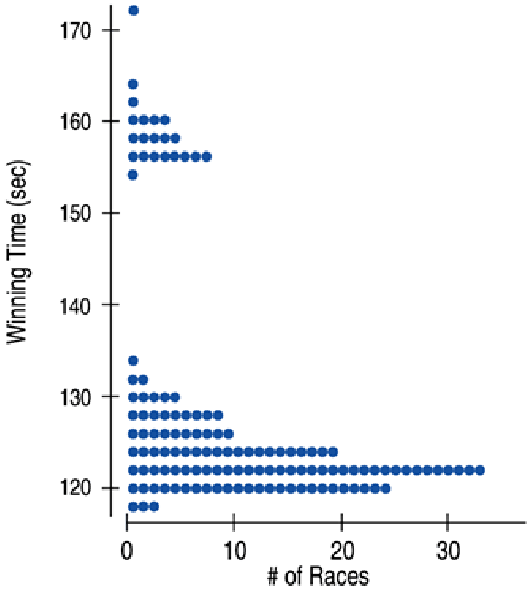
```
8 | 8
8 | 000044
7 | 6666
7 | 2222
6 | 8888
6 | 0444
5 | 6
```

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.

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.



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.

- When describing a distribution, make sure to always tell about three things:

What is the Shape of the Distribution?

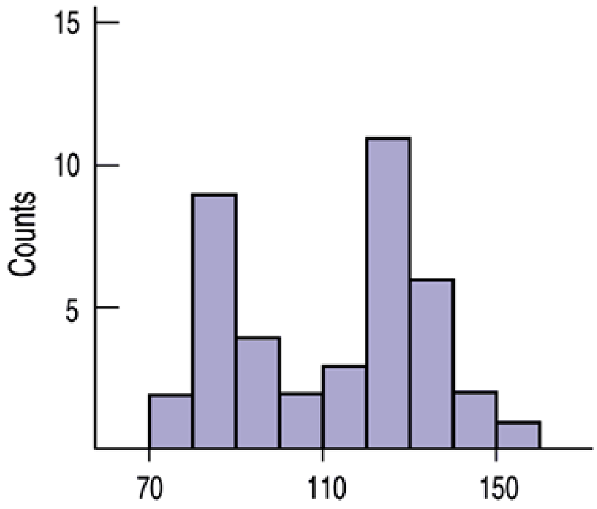
- Does the histogram have a single, central hump or several separated humps?
- Is the histogram symmetric?
- Do any unusual features stick out?

Humps

- Does the histogram have a single, central hump or several separated bumps?
- Humps in a histogram are called modes.

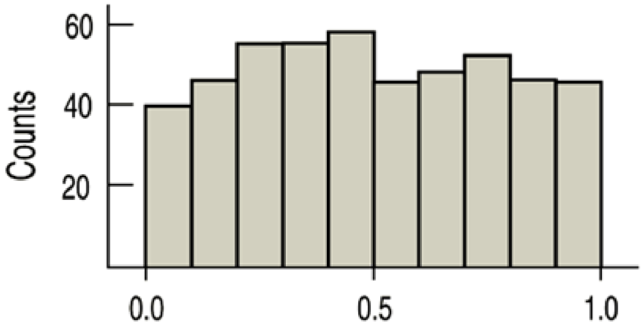
Humps (cont.)

- A bimodal histogram has two apparent peaks:



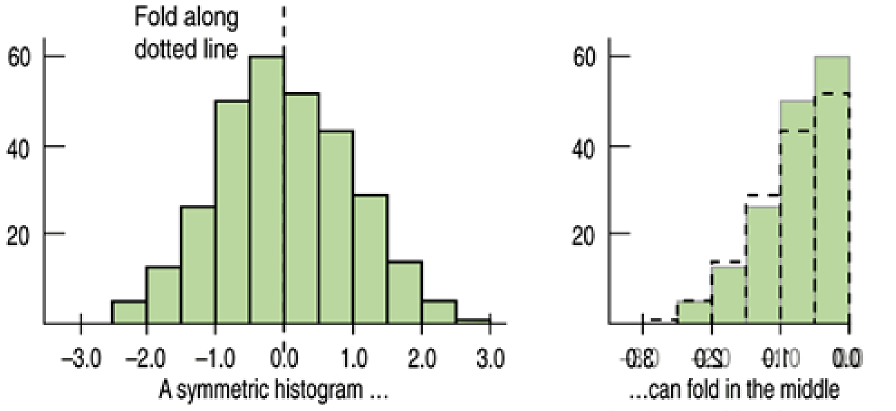
Humps (cont.)

- 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

- 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.

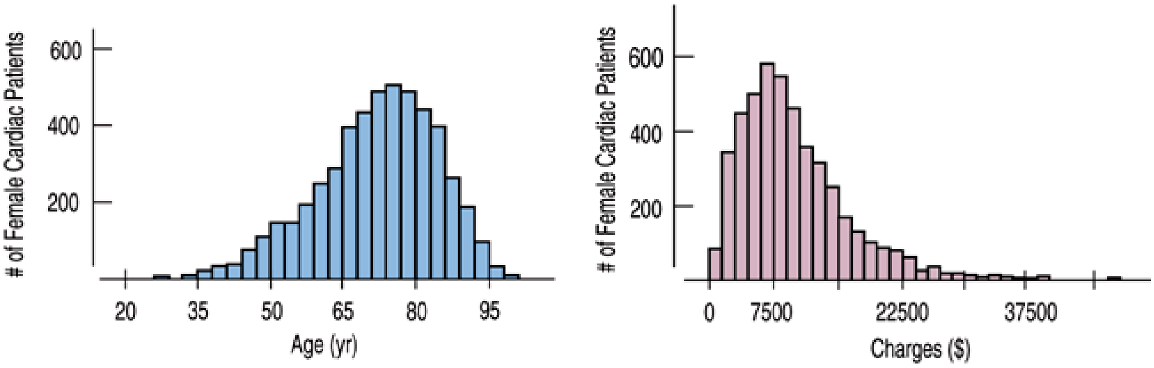


A symmetric histogram ...

...can fold in the middle so that the two sides almost match.

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.

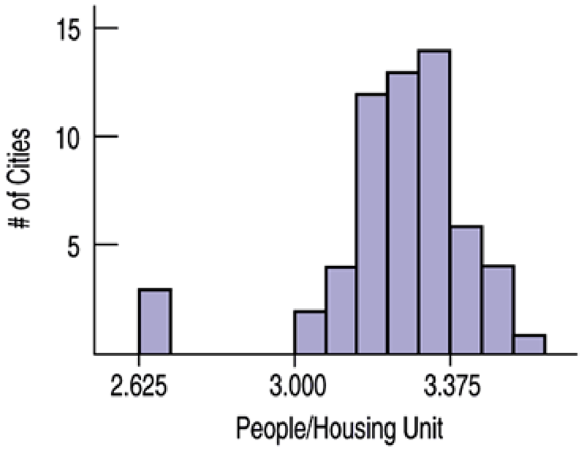


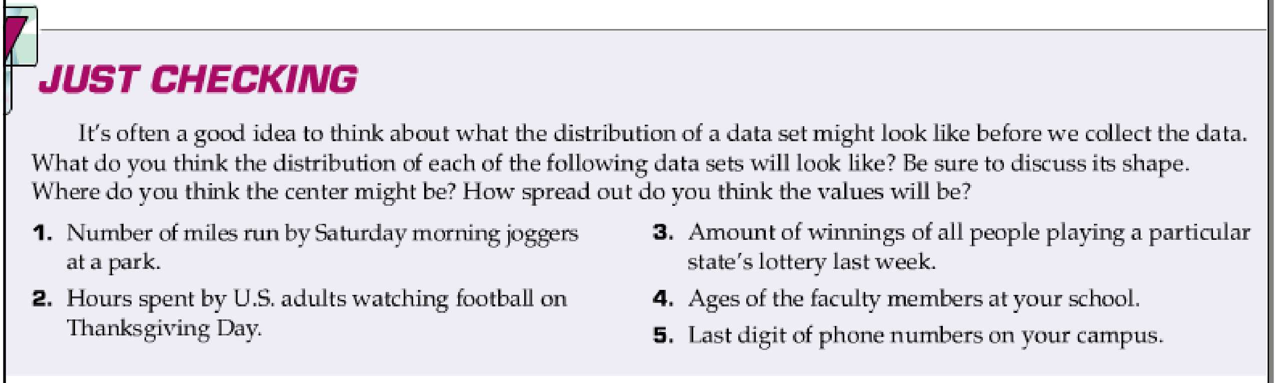
Anything Unusual?

- Do any unusual features stick out?
- Sometimes it’s the unusual features that tell us something interesting or exciting about the data.

Anything Unusual? (cont.)

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





JUST CHECKING

It's often a good idea to think about what the distribution of a data set might look like before we collect the data. What do you think the distribution of each of the following data sets will look like? Be sure to discuss its shape. Where do you think the center might be? How spread out do you think the values will be?

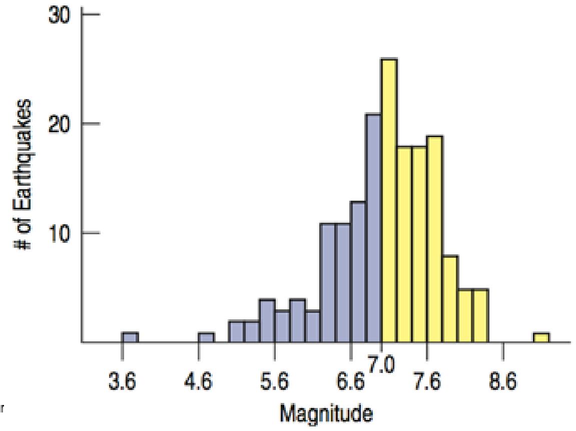
1. Number of miles run by Saturday morning joggers at a park.
2. Hours spent by U.S. adults watching football on Thanksgiving Day.
3. Amount of winnings of all people playing a particular state's lottery last week.
4. Ages of the faculty members at your school.
5. Last digit of phone numbers on your campus.

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



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.

Spread: Home on the Range

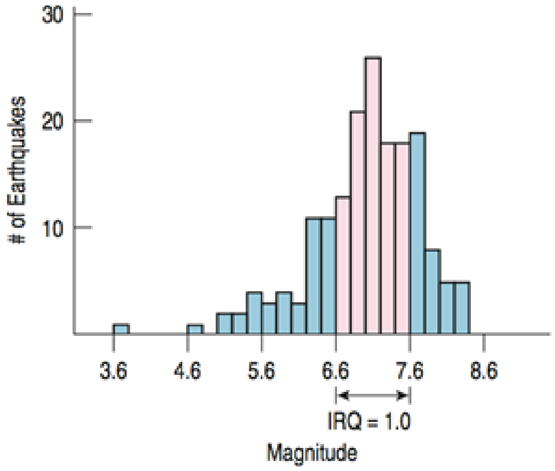
- The range of the data is the difference between the maximum and minimum values:
- Range = max – min
- A disadvantage of the range is that a single extreme value can make it very large and, thus, not representative of the data overall.

Spread: The Interquartile Range

- The interquartile range (IQR) lets us ignore extreme data values and concentrate on the middle of the data.
- To find the IQR, we first need to know what quartiles are…

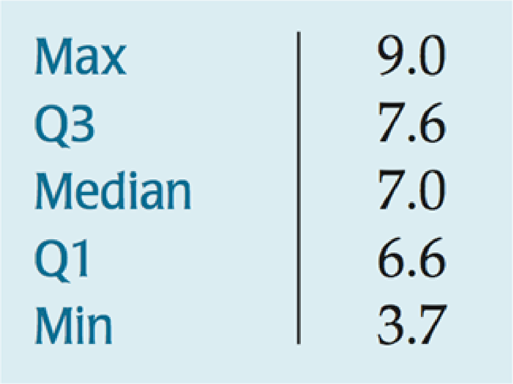
Spread: The Interquartile Range (cont.)

- The lower and upper quartiles are the 25th and 75th percentiles of the data, so…
- The IQR contains the middle 50% of the values of the distribution, as shown in figure:



5-Number Summary

- The 5-number summary of a distribution reports its median, quartiles, and extremes (maximum and minimum)
- The 5-number summary for the recent tsunami earthquake Magnitudes looks like this:



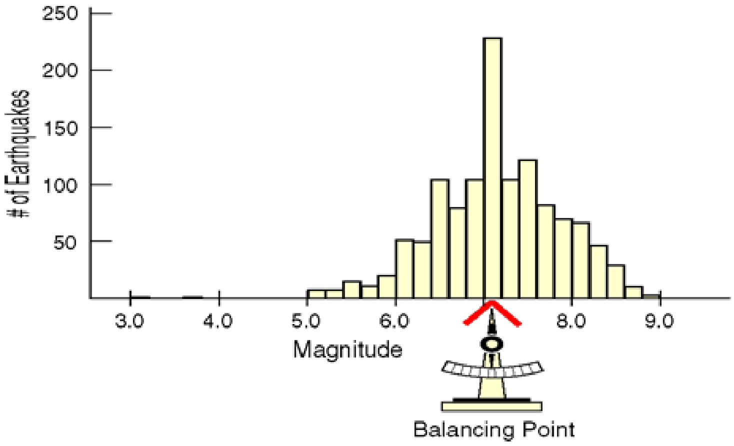
Max	9.0
Q3	7.6
Median	7.0
Q1	6.6
Min	3.7

Summarizing Symmetric Distributions -- The Mean

- When we have symmetric data, there is an alternative other than the median.
- If we want to calculate a number, we can average the data.
- We use the Greek letter sigma to mean "sum" and write:

Summarizing Symmetric Distributions -- The Mean (cont.)

- The mean feels like the center because it is the point where the histogram balances:



Mean or Median?

- Because the median considers only the order of values, it is resistant to values that are extraordinarily large or small; it simply notes that they are one of the “big ones” or “small ones” and ignores their distance from center.
- To choose between the mean and median, start by looking at the data. If the histogram is symmetric and there are no outliers, use the mean.
- However, if the histogram is skewed or with outliers, you are better off with the median.

What About Spread? The Standard Deviation

- A more powerful measure of spread than the IQR is the standard deviation, which takes into account how far each data value is from the mean.

A deviation is the distance that a data value is from the mean.

- Since adding all deviations together would total zero, we square each deviation and find an average of sorts for the deviations.

What About Spread? The Standard Deviation (cont.)

- The variance, notated by s^2, is found by summing the squared deviations and (almost) averaging them:

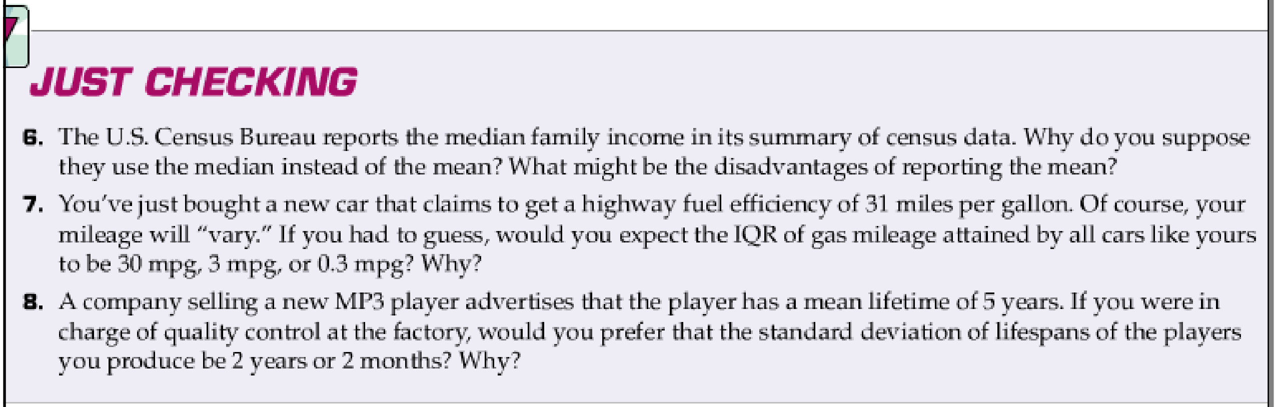
- The variance will play a role later in our study, but it is problematic as a measure of spread—it is measured in squared units!

What About Spread? The Standard Deviation (cont.)

- The standard deviation, s, is just the square root of the variance and is measured in the same units as the original data.

Thinking About Variation

- Since Statistics is about variation, spread is an important fundamental concept of Statistics.
- Measures of spread help us talk about what we don’t know.



JUST CHECKING

6. The U.S. Census Bureau reports the median family income in its summary of census data. Why do you suppose they use the median instead of the mean? What might be the disadvantages of reporting the mean?

7. You've just bought a new car that claims to get a highway fuel efficiency of 31 miles per gallon. Of course, your mileage will "vary." If you had to guess, would you expect the IQR of gas mileage attained by all cars like yours to be 30 mpg, 3 mpg, or 0.3 mpg? Why?

8. A company selling a new MP3 player advertises that the player has a mean lifetime of 5 years. If you were in charge of quality control at the factory, would you prefer that the standard deviation of lifespans of the players you produce be 2 years or 2 months? Why?

Tell -- Draw a Picture

- When telling about quantitative variables, start by making a histogram or stem-and-leaf display and discuss the shape of the distribution.

Tell -- Shape, Center, and Spread

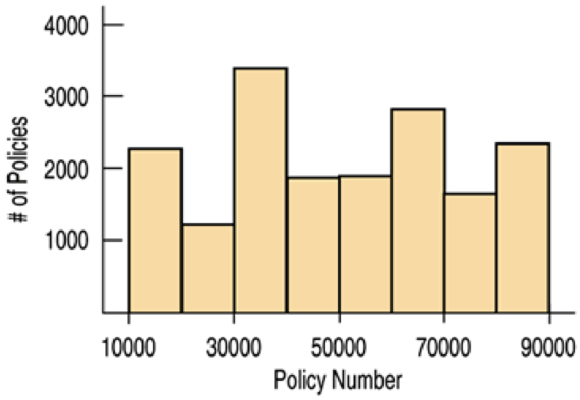
- Next, always report the shape of its distribution, along with a center and a spread.
- If the shape is skewed,
- If the shape is symmetric,

Tell -- What About Unusual Features?

- If there are multiple modes, try to understand why. If you identify a reason for the separate modes, it may be good to split the data into two groups.
- If there are any clear outliers and you are reporting the mean and standard deviation, report them with the outliers present and with the outliers removed. The differences may be quite revealing.
- Note: The median and IQR are not likely to be affected by the outliers.

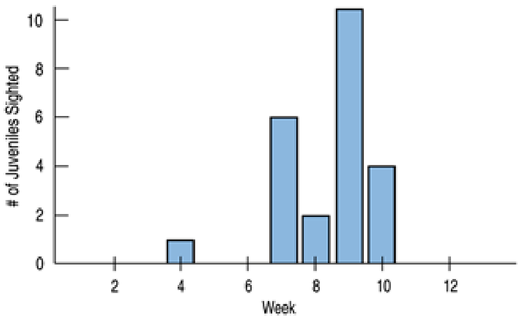
What Can Go Wrong?

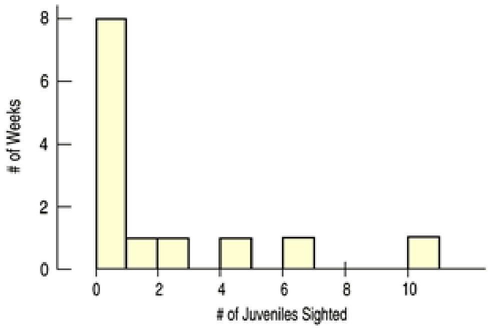
- Don’t make a histogram of a categorical variable—bar charts or pie charts should be used for categorical data.
- Don’t look for shape, center, and spread of a bar chart.



What Can Go Wrong? (cont.)

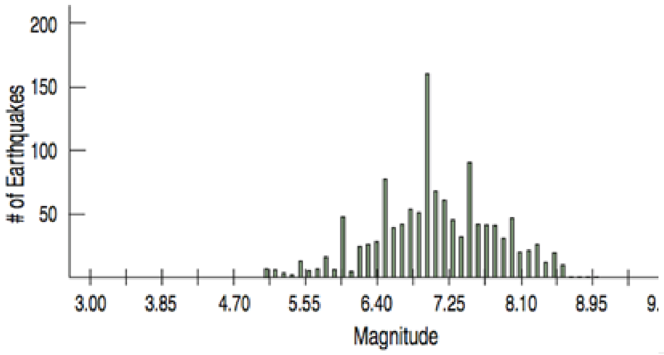
- Don’t use bars in every display—save them for histograms and bar charts.
- Below is a badly drawn plot and the proper histogram for the number of juvenile bald eagles sighted in a collection of weeks:

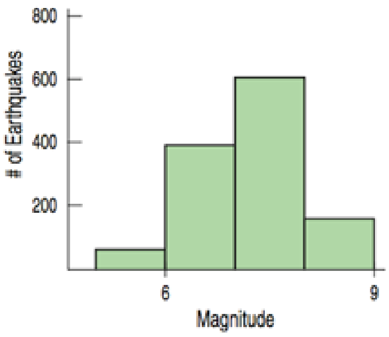




What Can Go Wrong? (cont.)

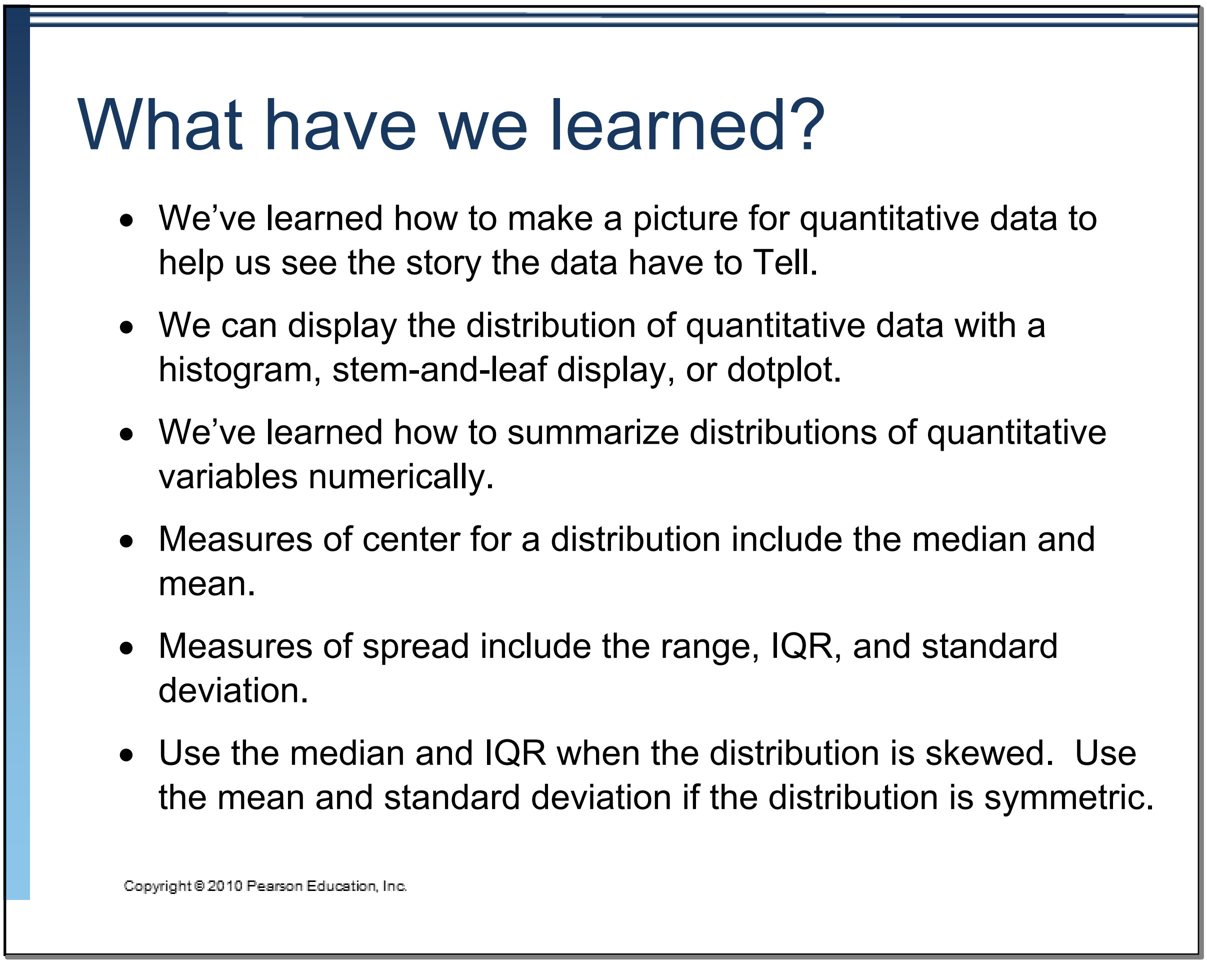
- Choose a bin width appropriate to the data.
- Changing the bin width changes the appearance of the histogram:





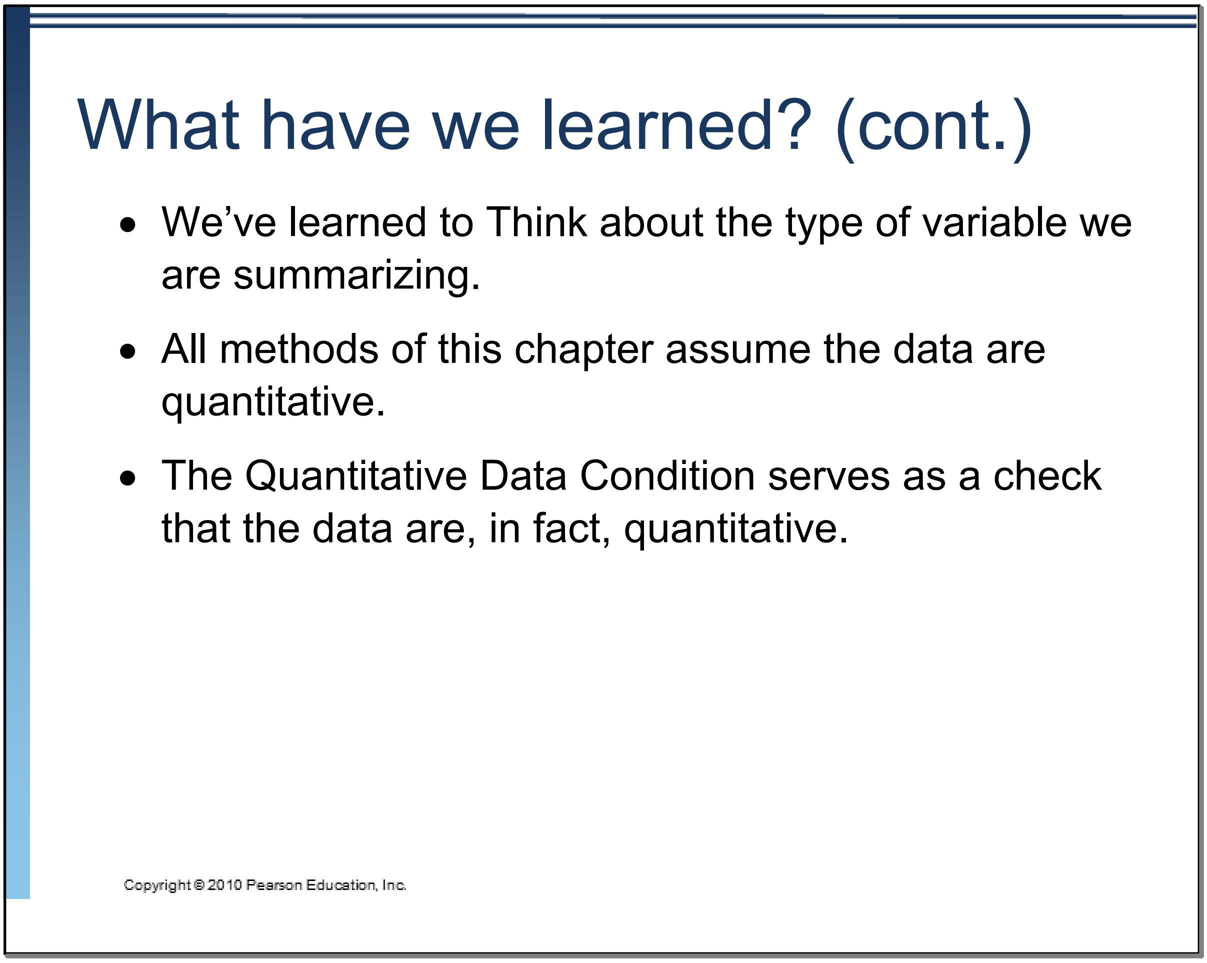
What Can Go Wrong? (cont.)

- Don’t forget to do a reality check – don’t let the calculator do the thinking for you.
- Don’t forget to sort the values before finding the median or percentiles.
- Don’t worry about small differences when using different methods.
- Don’t compute numerical summaries of a categorical variable.
- Don’t report too many decimal places.
- Don’t round in the middle of a calculation.
- Watch out for multiple modes
- Beware of outliers
- Make a picture … make a picture . . . make a picture !!!



What have we learned?

- We’ve learned how to make a picture for quantitative data to help us see the story the data have to Tell.
- We can display the distribution of quantitative data with a histogram, stem-and-leaf display, or dotplot.
- We’ve learned how to summarize distributions of quantitative variables numerically.
- Measures of center for a distribution include the median and mean.
- Measures of spread include the range, IQR, and standard deviation.
- Use the median and IQR when the distribution is skewed. Use the mean and standard deviation if the distribution is symmetric.



What have we learned? (cont.)

- We’ve learned to Think about the type of variable we are summarizing.
- All methods of this chapter assume the data are quantitative.
- The Quantitative Data Condition serves as a check that the data are, in fact, quantitative.

Ex. 1: How good was the 2004 U.S. women's soccer team? They put on an impressive showing en route to winning the gold medal at the 2004 Olympics. Here are data on the number of goals scored by the team in 34 games played during the 2004 season. Make a histogram or dotplot for the data.

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

Ex. 2: How many pairs of shoes does a typical teenager have? To find out, a group of AP Stats students conducted a survey. They selected a random sample of 20 female students from their school. Then they recorded the number of pairs of shoes that each respondent reported having. Make a stemplot for the data below:

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

Ex. 3: Refer to the data on travel times for a sample of 15 North Carolinians below. Find the mean travel time for the 15 worker. Are there any outliers? If so, find the mean of the travel time excluding the outlier(s) and report what you discover.

0	5
1	000025
2	005
3	00
4	00
5	
6	0

Ex. 4: People say that it takes a long time to get to work in New York State due to the heavy traffic near big cities. What do the data say? Here are the travel times in minutes of 20 randomly chosen New York workers:

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

Ex. 5: The heights (in inches) of five starters on a basketball team are 67, 72, 76, 76, and 84.

1. Find and interpret the mean.
2. Make a table that shows, for each value, its deviation from the mean, and its squared deviation from the mean.
3. Show how to calculate the variance and standard deviation from the values in your table.
4. Interpret the meaning of the standard deviation in this setting.

1. Mean:

2. Table:

Observation	Deviation	Squared Deviation
●	●	●
●	●	●
●	●	●
●	●	●
●	●	●
●	●	●

3. Variance:

   Standard deviation:

4. Interpretation: