Stat 201: Statistics I
Week 2

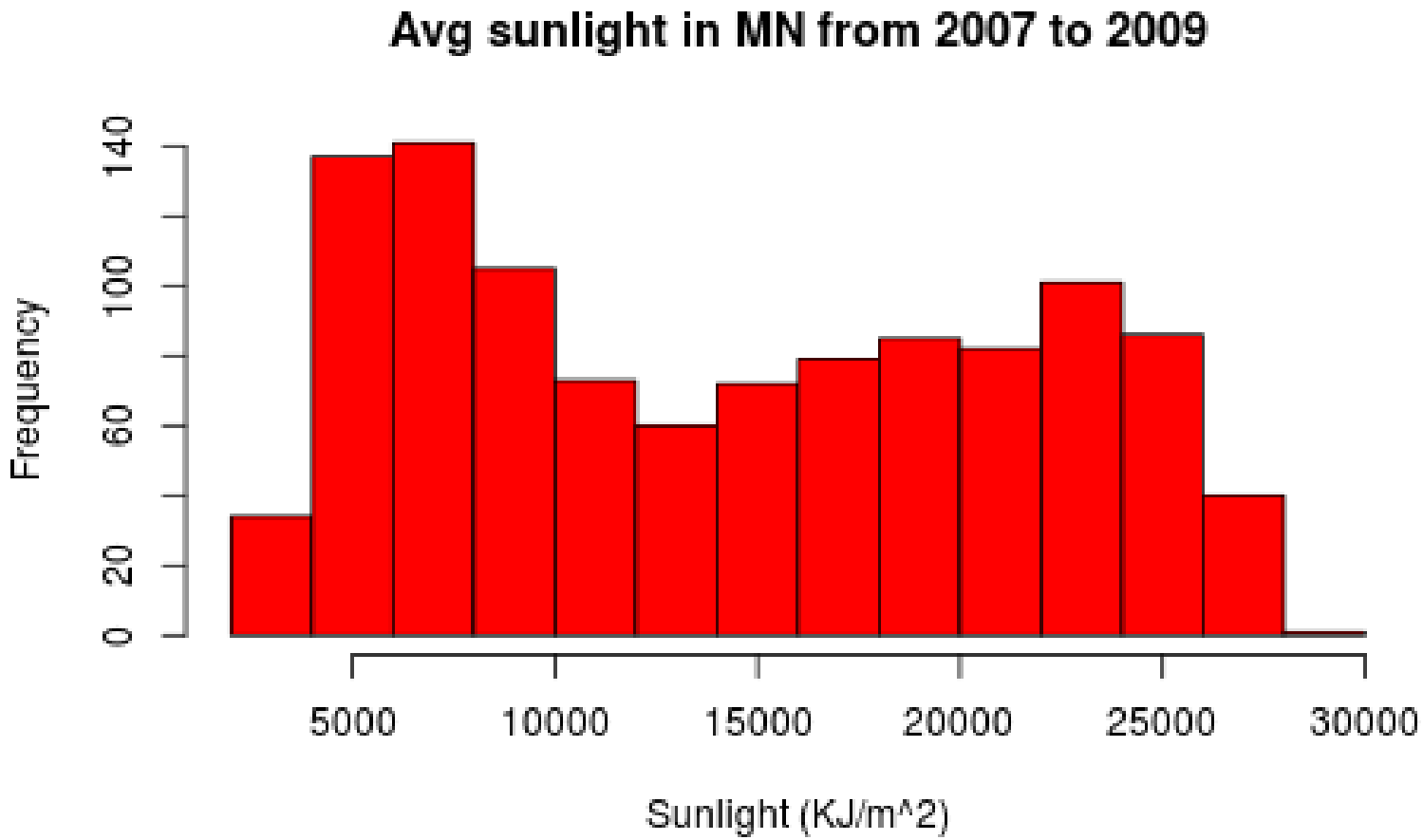
May 22, 2017

Chapter 2
Summarizing and Graphing Data

Section 2.3
Histograms

Histograms

A **histogram** is a graphical representation of a frequency distribution of quantitative data. This allows the distribution of the data to be more easily visualized.



Properties of histograms

- A graph of bars of equal width drawn adjacent to each other.
- The horizontal scale (x-axis) represents values of the quantitative data. Each bar represents a class, or range of values, from a frequency table.
- The vertical scale (y-axis) represents frequency (counts), or proportions (relative frequency) or percentages (percentage frequency).
- The number of bars is largely an aesthetic choice. There should be enough bars to adequately show the shape of the distribution, but too many can make a "busy" graph that's hard to read. Most software will automatically choose the number of bars.

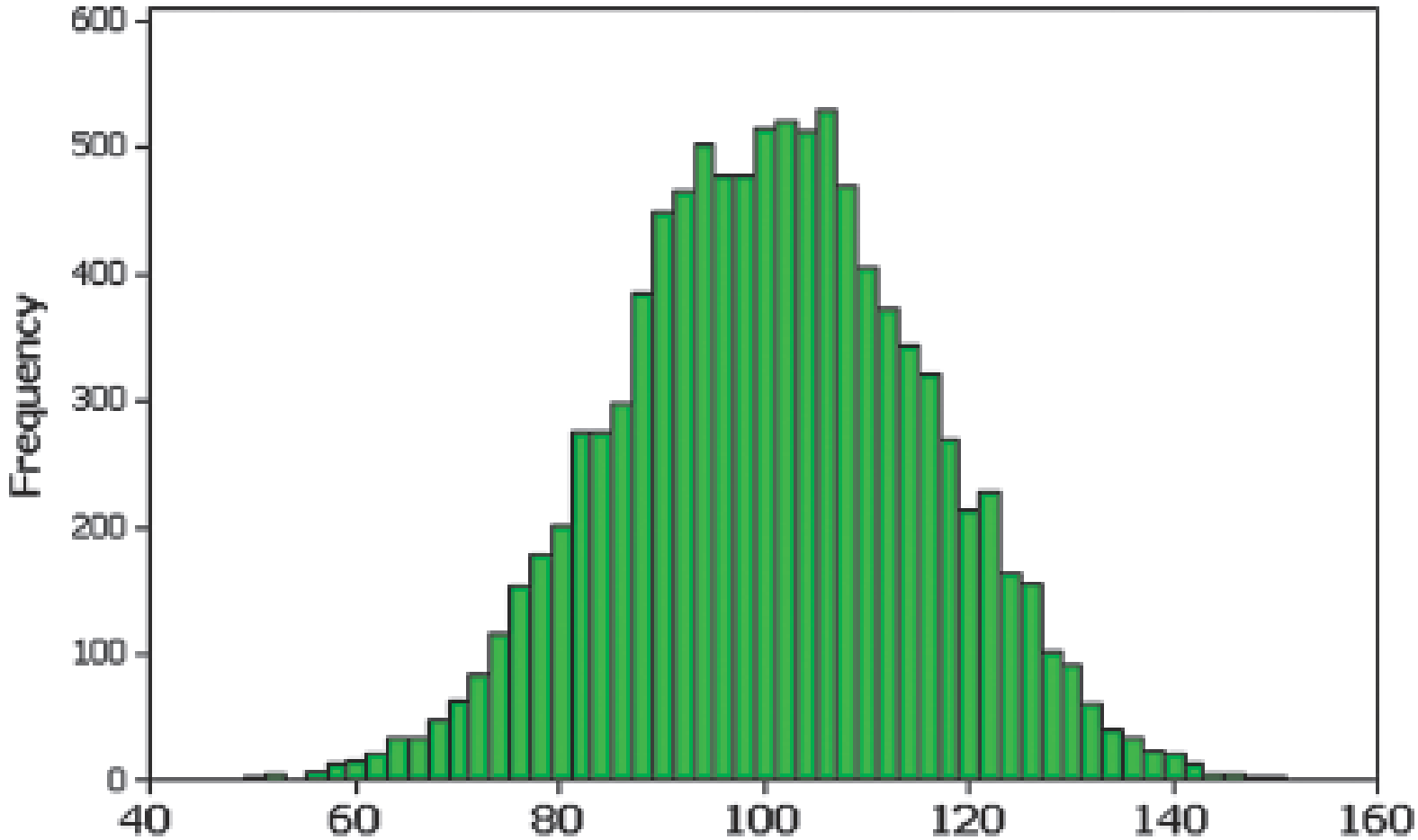
Histograms and normal distributions

Remember: A **normal distribution** can be identified from a frequency table that has the following characteristics:

- The frequencies start low, increase to a high point and then decrease to low frequencies at the end
- The frequencies are approximately symmetric around the high point.

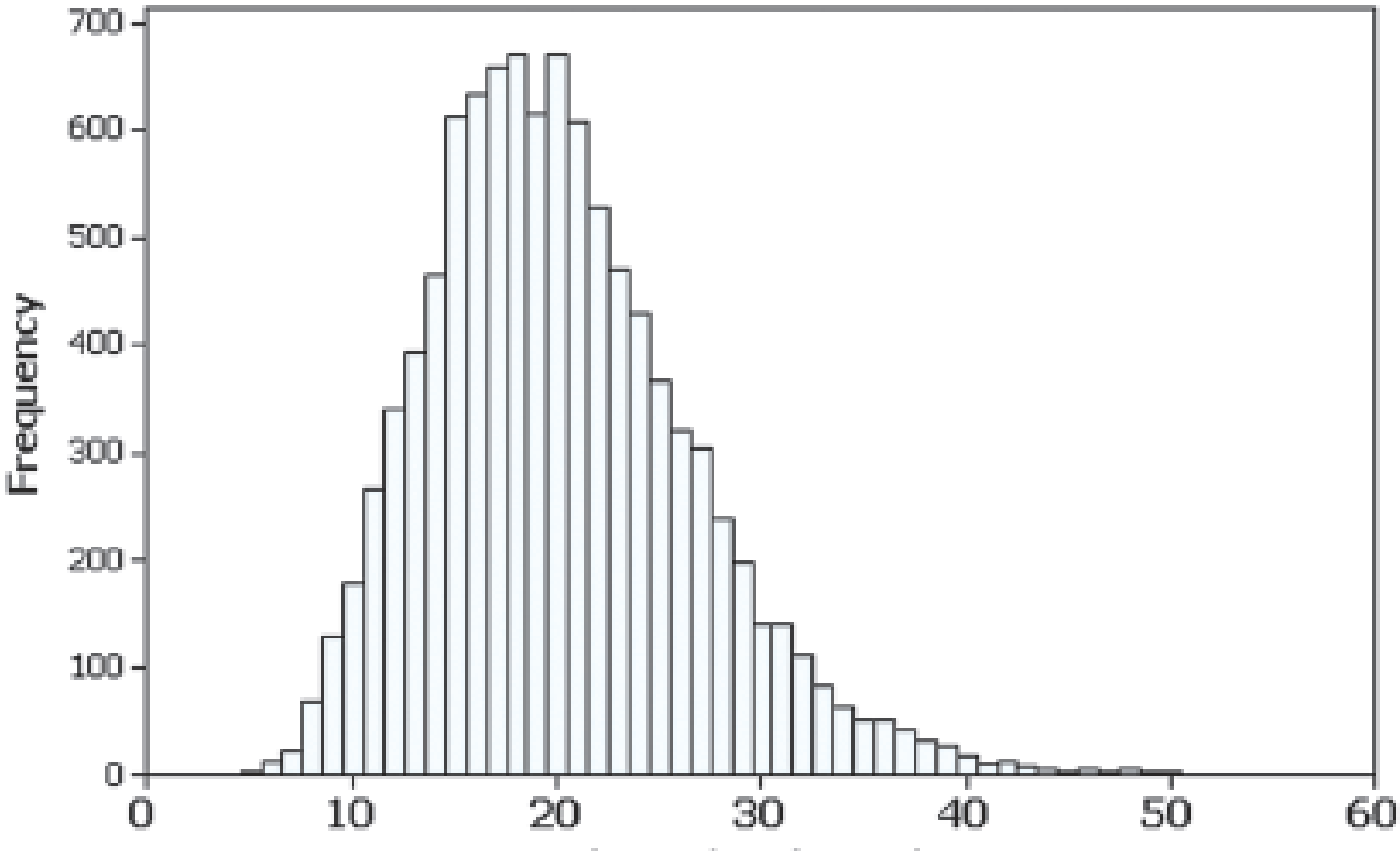
Graphically, normal distributions are commonly known as "bell curves". Histograms can be used to recognize when data follows a normal distribution.

Histograms and normal distributions, examples



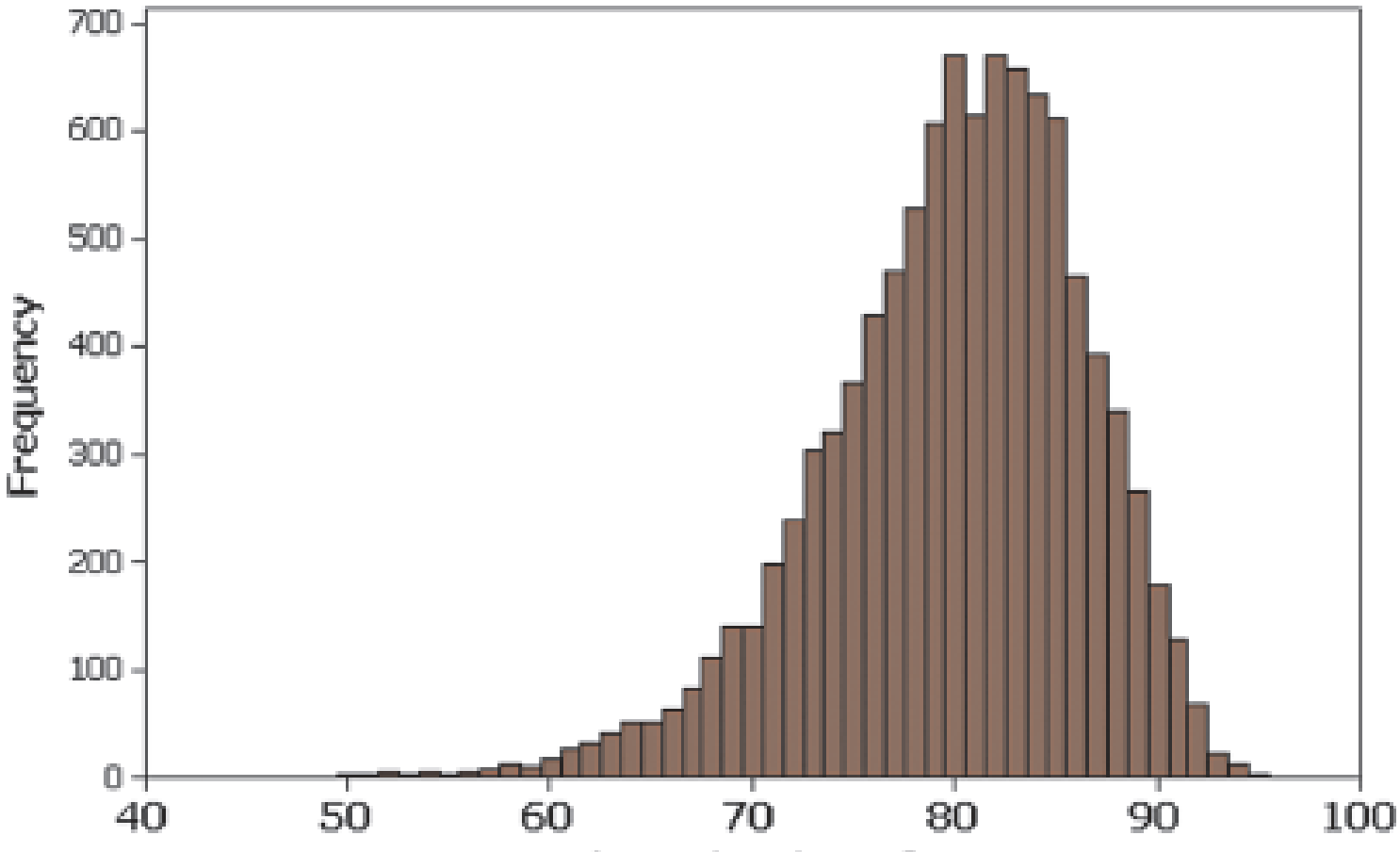
Normal

Histograms and normal distributions, examples



Right skewed

Histograms and normal distributions, examples



Left skewed

Section 2.4
Graphs that Enlighten and Graphs that Deceive

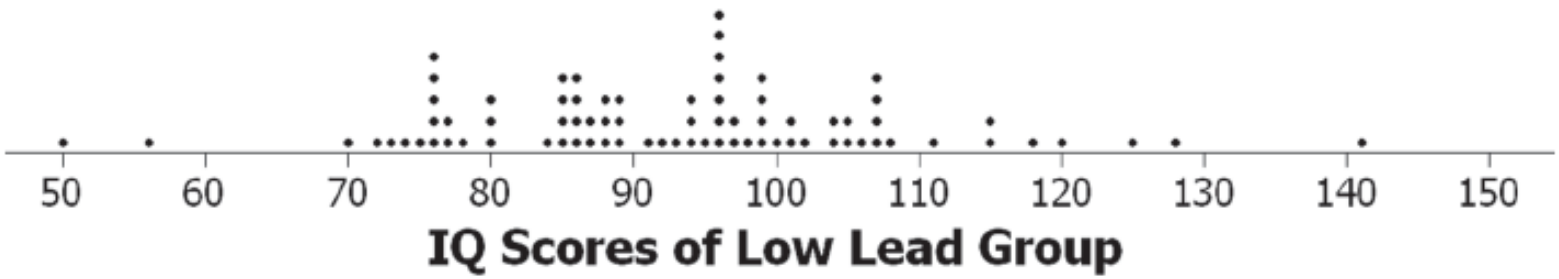
Types of graphs

There are many types of graphs. Deciding which to use depends on the type of data involved and the message to be delivered.

Types of graphs: dotplots

A **dotplot** is similar to a histogram.

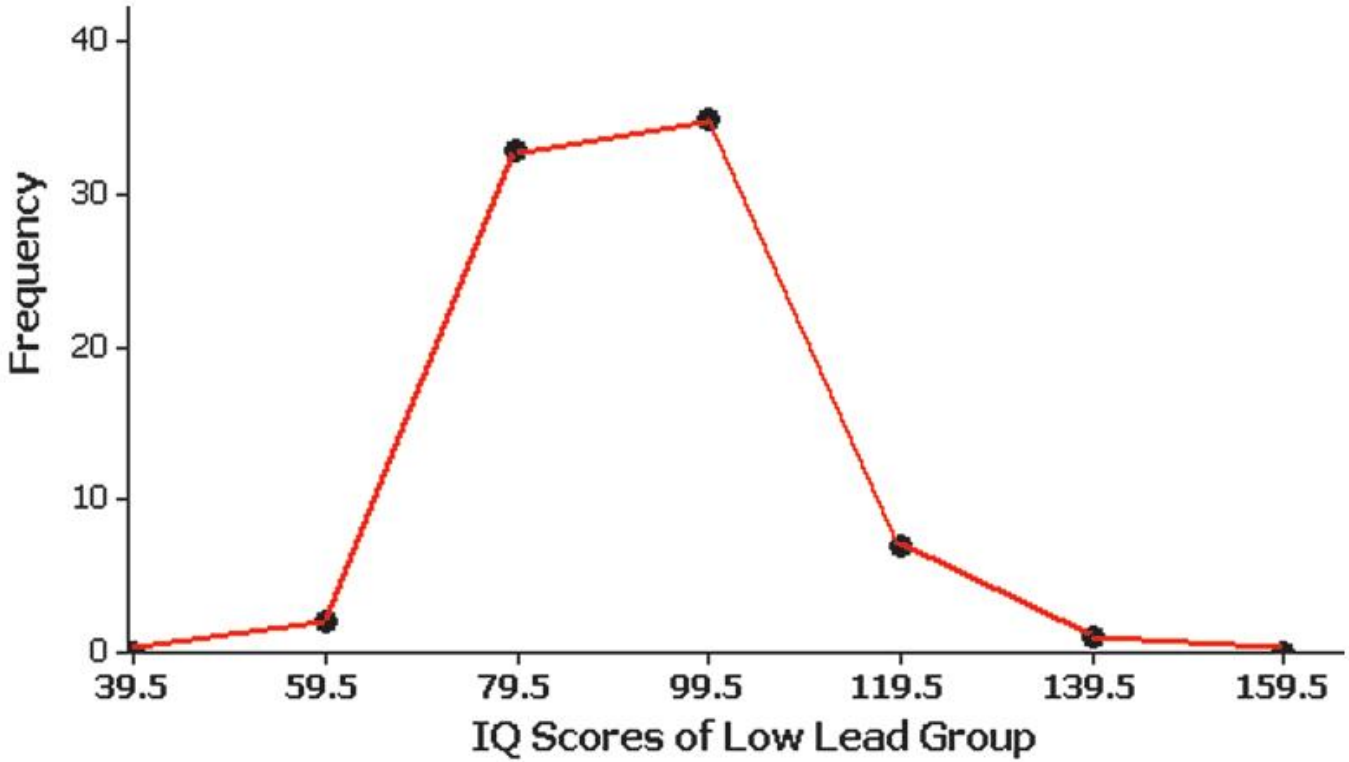
- The x-axis represents values of the quantitative data
- Instead of bars, a dot is placed for each instances of a value



Types of graphs: frequency polygon

A **frequency polygon** is similar to a histogram.

- Instead of bars, a single dot is drawn above the midpoint of each class at a height representing frequency.
- Lines are drawn between the points.



Types of graphs: stem-and-leaf plots

A **stem-and-leaf plot** is also used display frequencies of quantitative data

- Each numeric value is separated into two parts, the leftmost digits (the stem) and the last digit (the leaf). For example, 142 $\Rightarrow$ 14 and 2.
- Each stem is arranged vertically on the left side of the graph.
- Every leaf belonging to a stem is listed to the right, in numeric order.

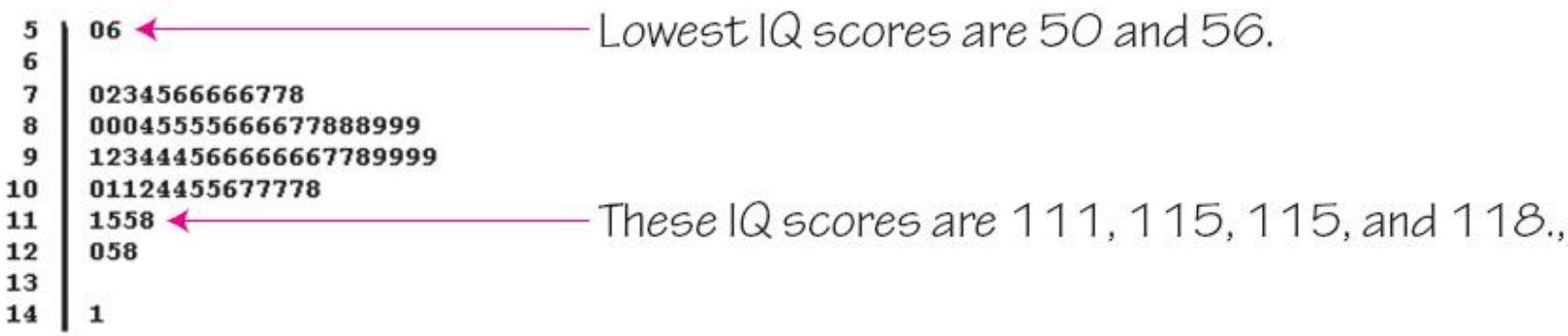
Example

Value	$\Rightarrow$	Stem	Leaf
142		14	2
146		14	6
138		13	8
143		14	3

Stem-and-leaf plot

13	8
14	2 3 6

Stem-and-leaf plot, example

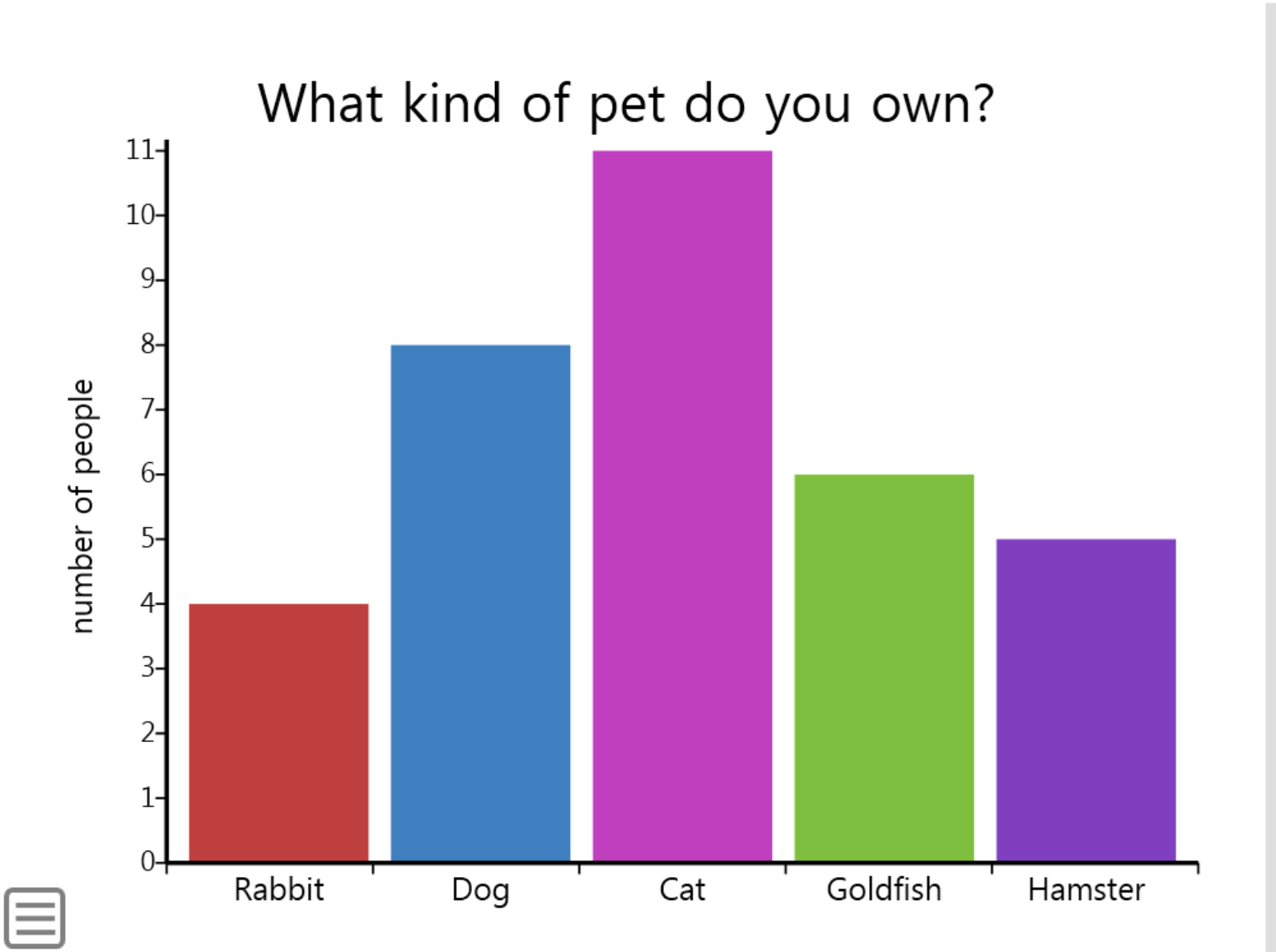


Types of graphs: bar graph

A **bar graph** displays frequencies of categorical data.

- The horizontal scale (x-axis) represents values of the categorical data.
- The vertical scale (y-axis) represents frequencies (or proportions or percentages).
- Often, but not always, bars are drawn with a gap between values.

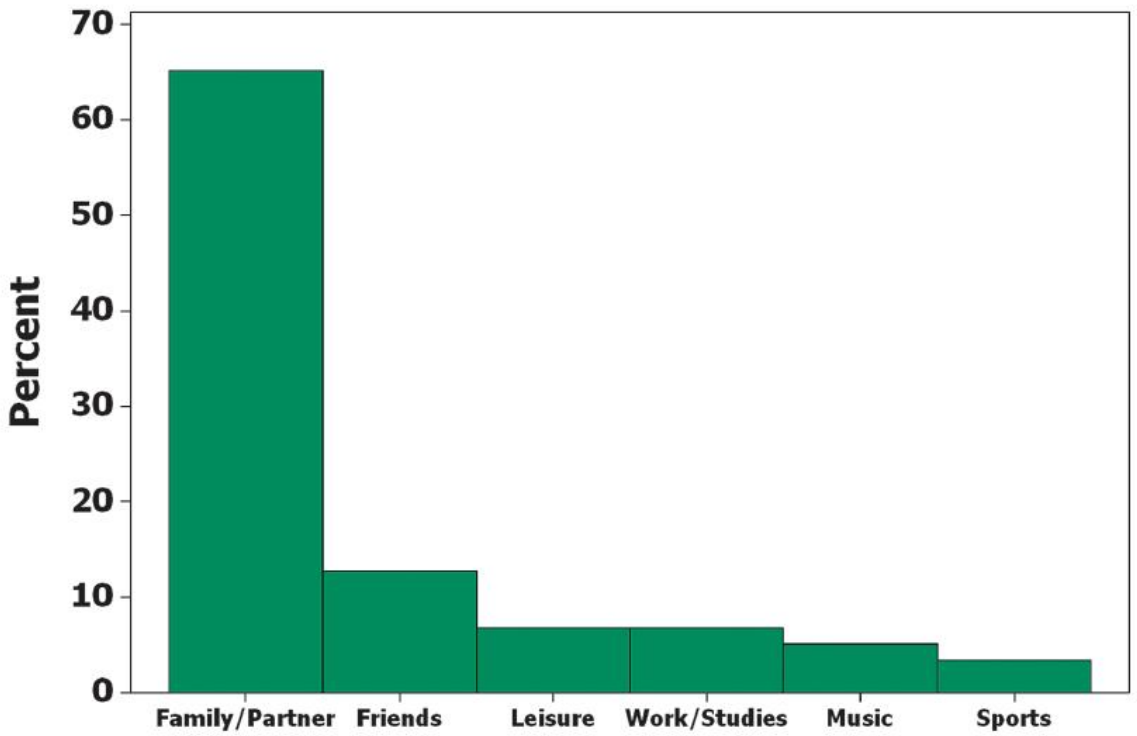
Bar graph, example



Types of graphs: Pareto charts

A **Pareto chart** is very similar to a bar graph, except the bars are arranged from most frequent to least, left to right.

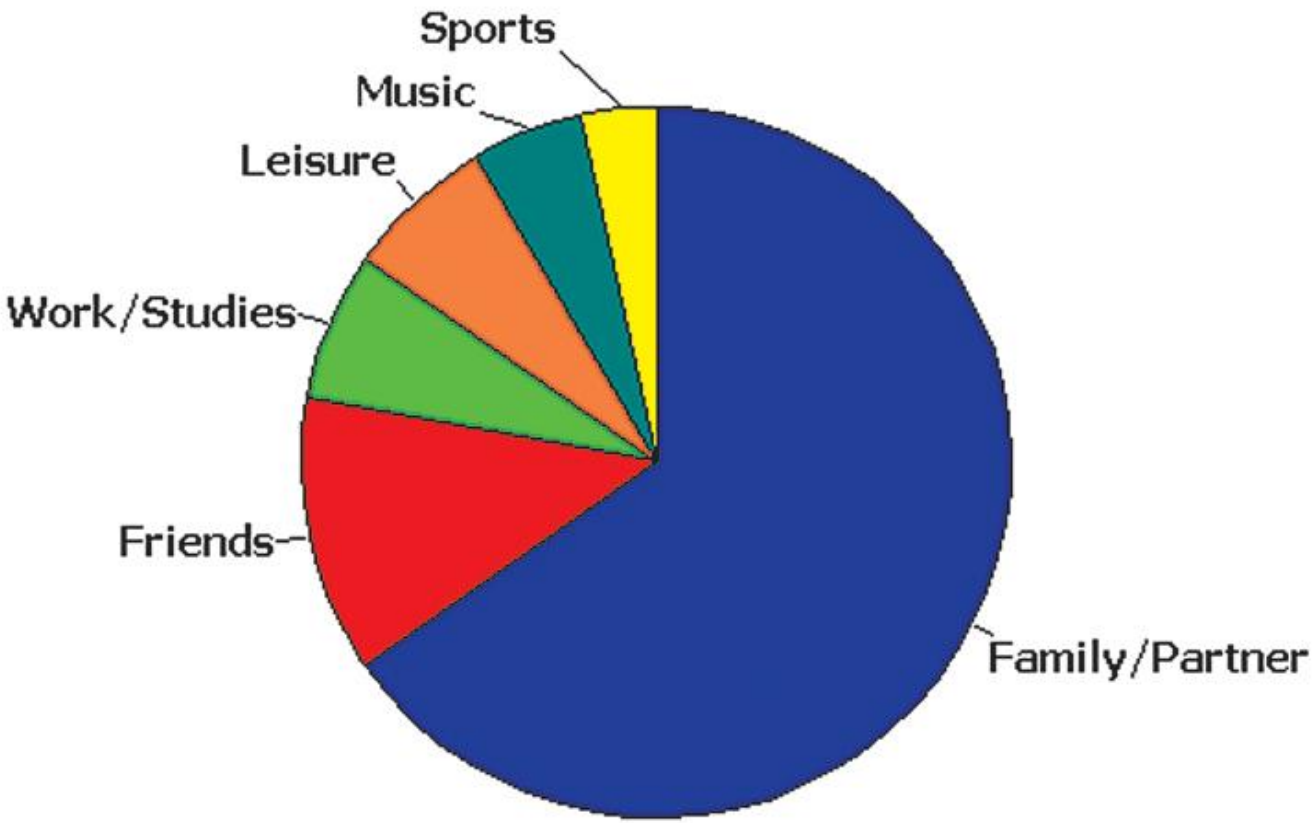
- Can be confusing if used with ordinal data.



Pareto Chart: What Contributes Most to Happiness?

Types of graphs: pie charts

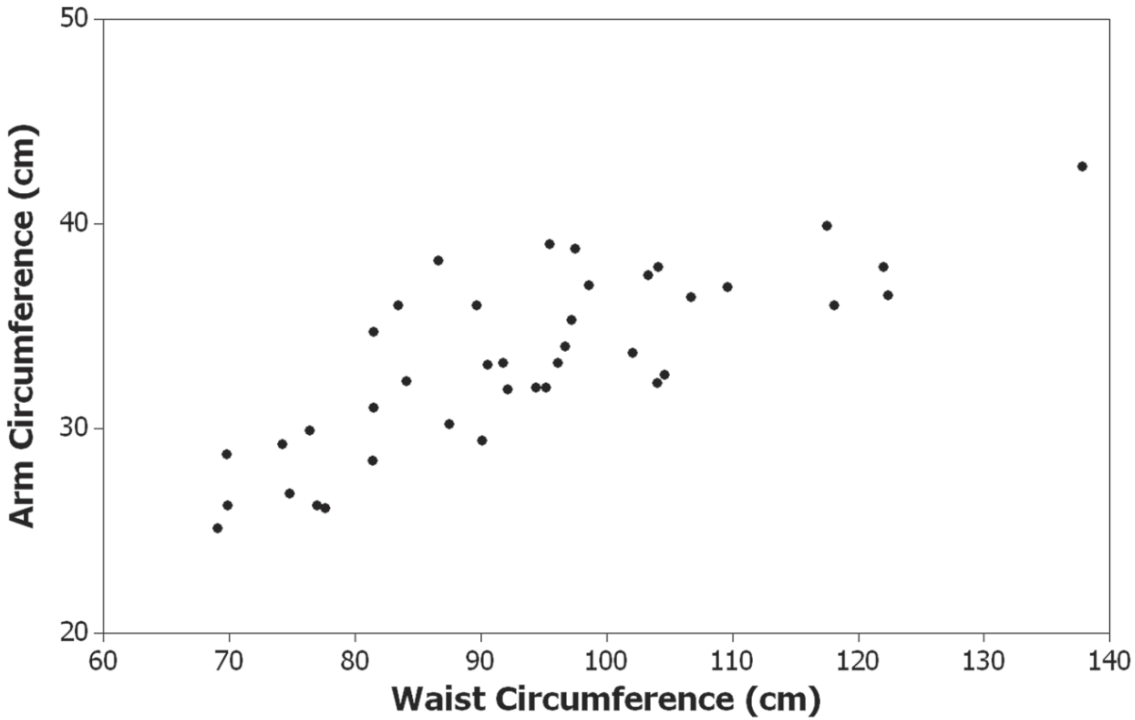
A **pie chart** displays relative frequencies of categorical data as "slices" of a whole circle. The "slices" must be labelled or distinguished by color.



Types of graphs: scatterplots

A **scatterplot** displays the relationship between paired quantitative variables.

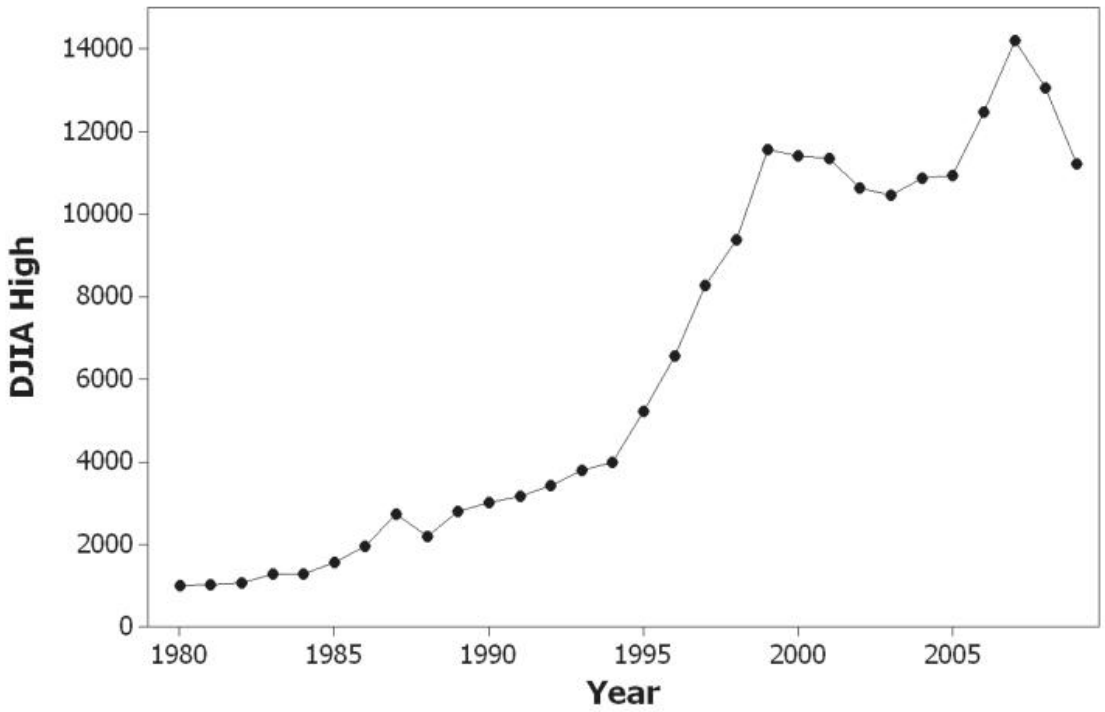
- The x-axis represents one variable and the y-axis the other.
- A dot (or other symbol) for each data pair is placed at the appropriate x and y values.



Types of graphs: time series

A graph of paired quantitative data where one variable represents time is called a **time series**. It is much like a scatterplot, except...

- The x-axis always represents the time variable.
- Often a line is drawn between the points.



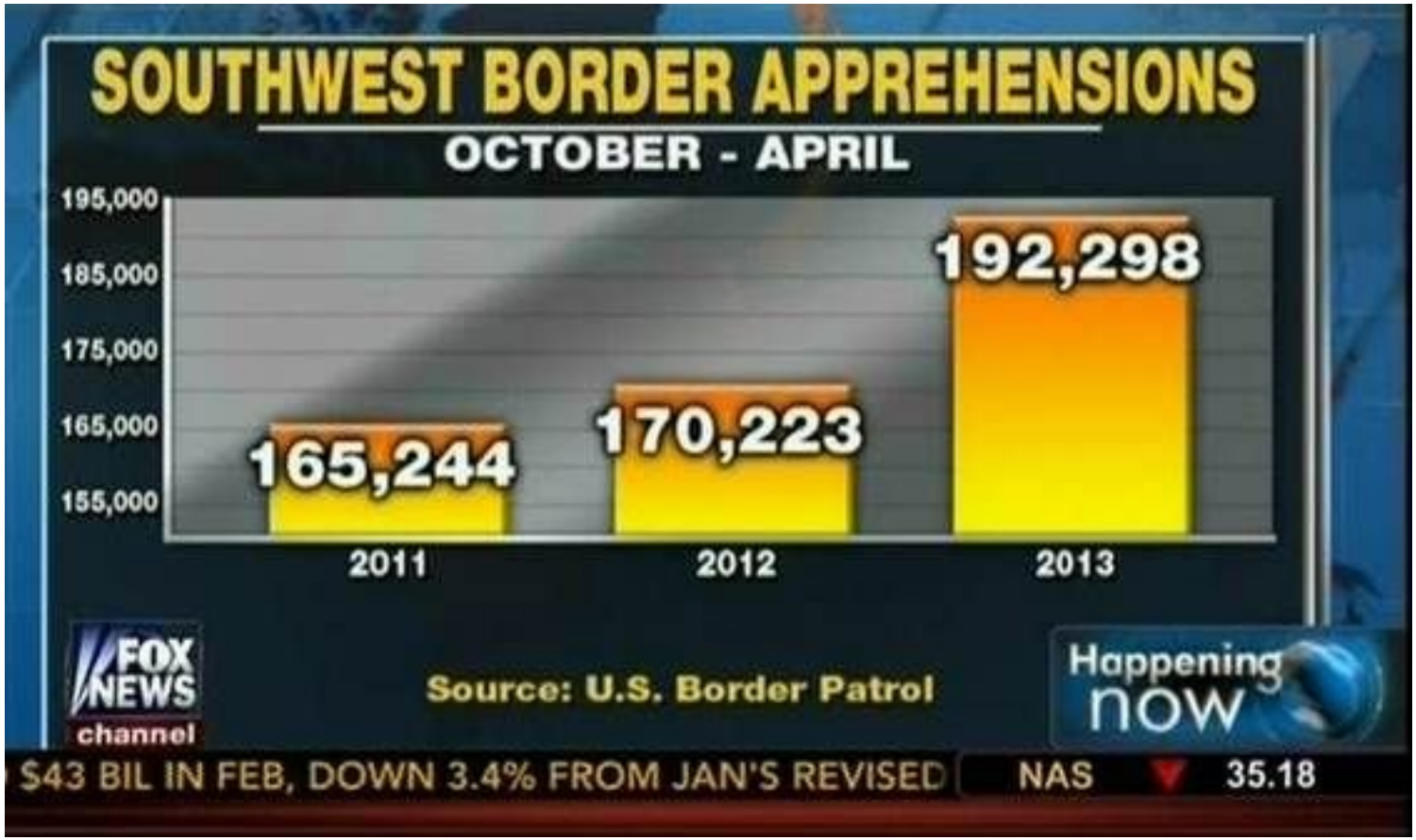
Graphs that deceive

There are two types of bad graphs:

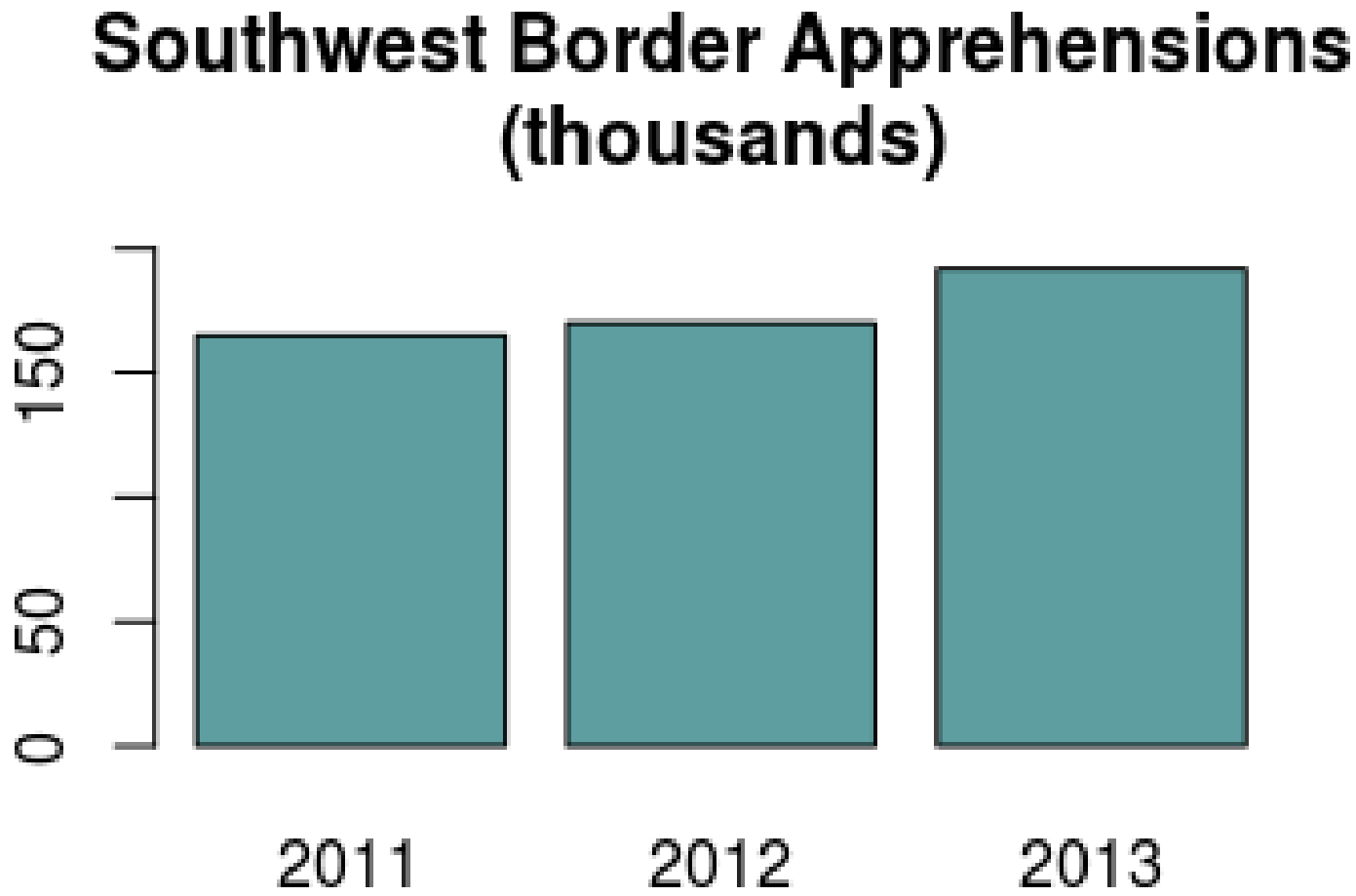
- Sometimes a graph is factually incorrect, whether because of errors in the data or a mistake in creating the graph. This is often difficult to detect without access to the original data.
- Sometimes graphs are technically correct, but designed to give a false impression of the data. Part of being a critical consumer of statistics is learning to recognize these misleading graphs.

Misleading graphs: non-zero axis

A **non-zero axis** is when one of the axis has a scale which does not include zero. This can make the relative sizes of the graph items to be distorted, especially in histograms or bar graphs.



Non-zero axis, fixed



Misleading graphs: pictographs

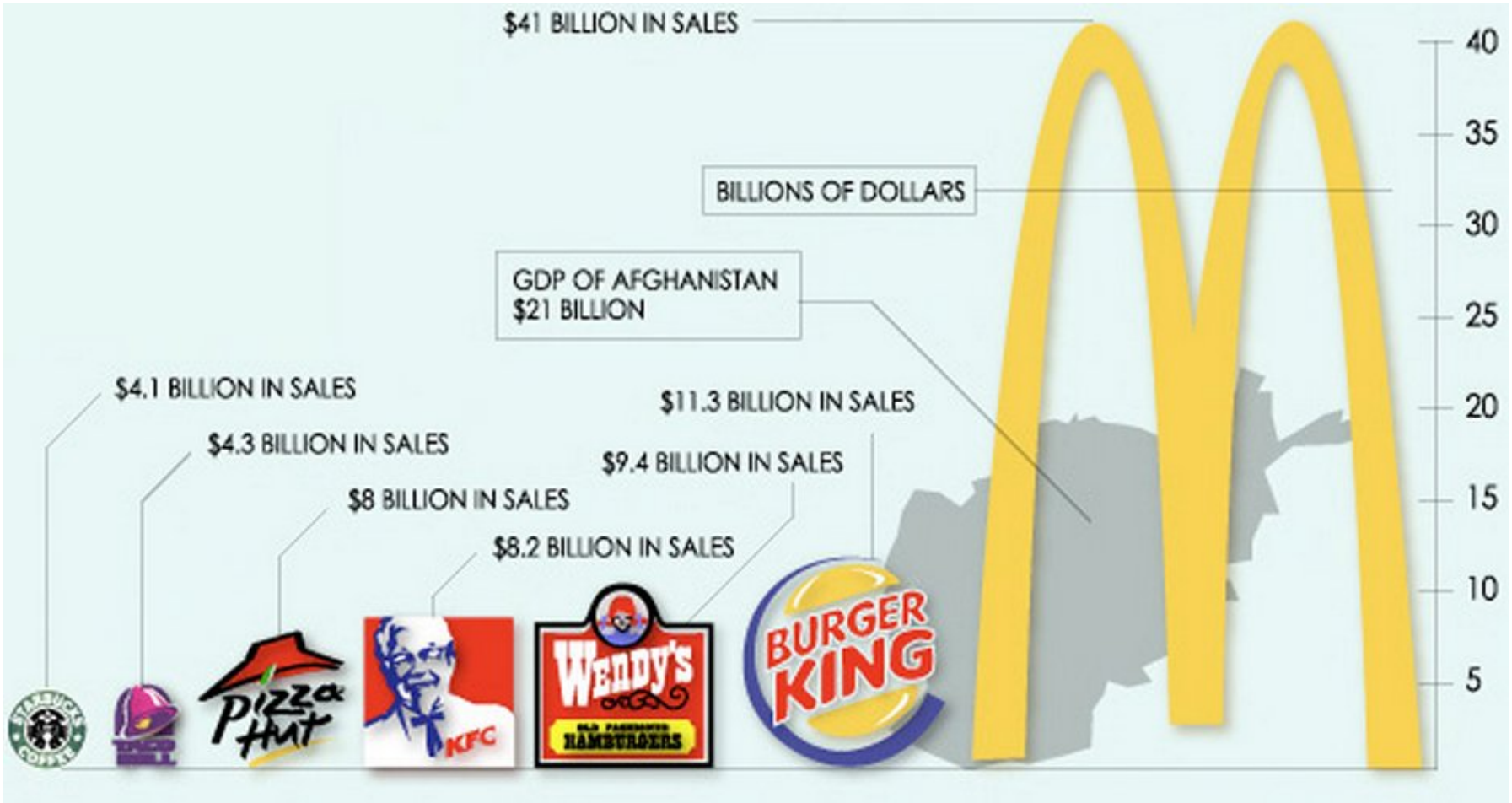
A **pictograph** uses pictures or 3D objects to represent size, rather than simple bars or points. This can also distort relative sizes.

Example

Suppose we wanted to graph the difference in sales between two oil companies, one of which is has twice the sales as the other. If we created a pictograph, we would draw the height of the larger sales twice as tall as the other.

- If we used a pictures, such as a company logos, the larger would have 4 times the area.
- If we used a 3D object, such as an oil barrel, the larger would have 8 times the volume.

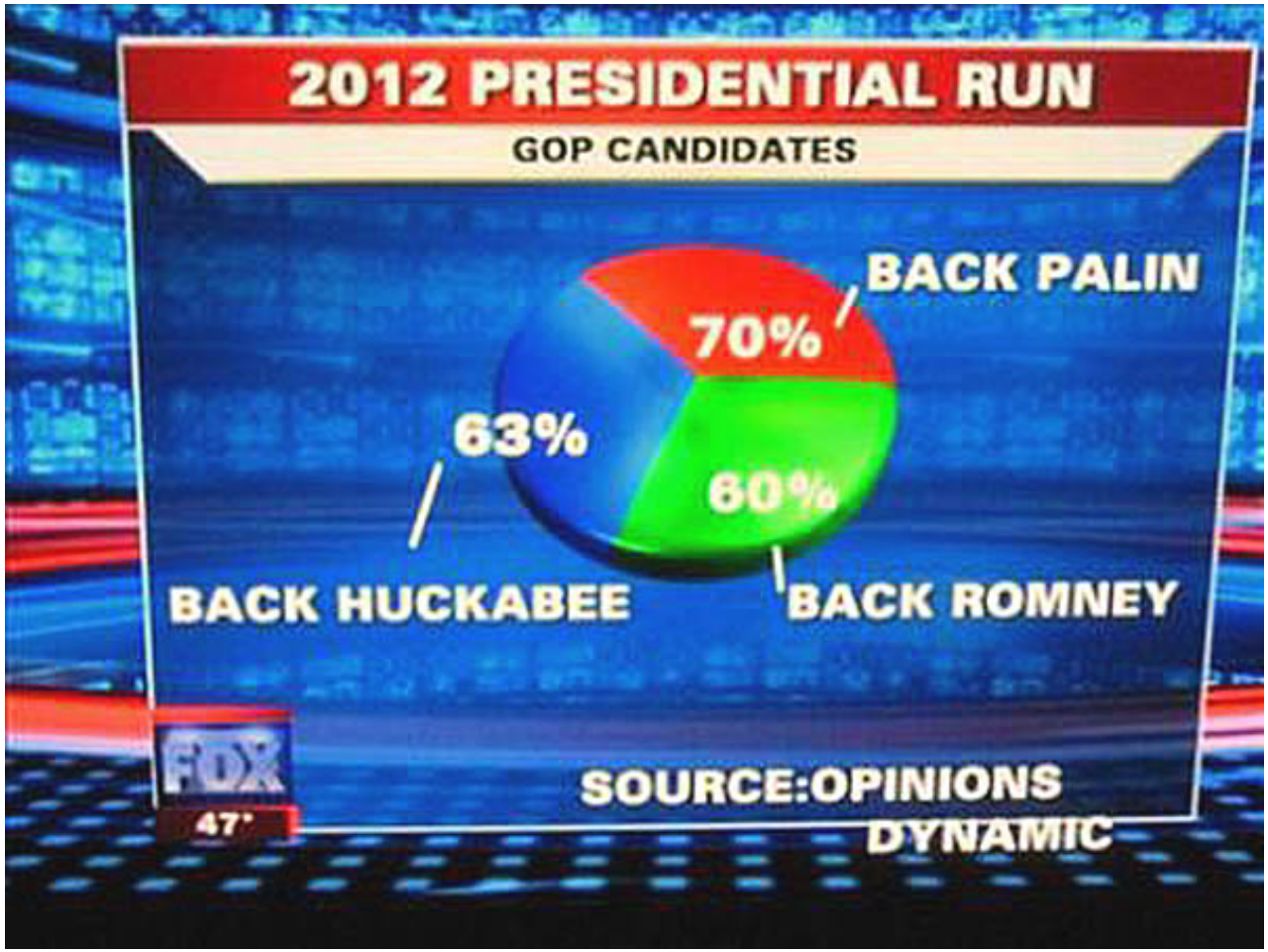
Pictograph, example



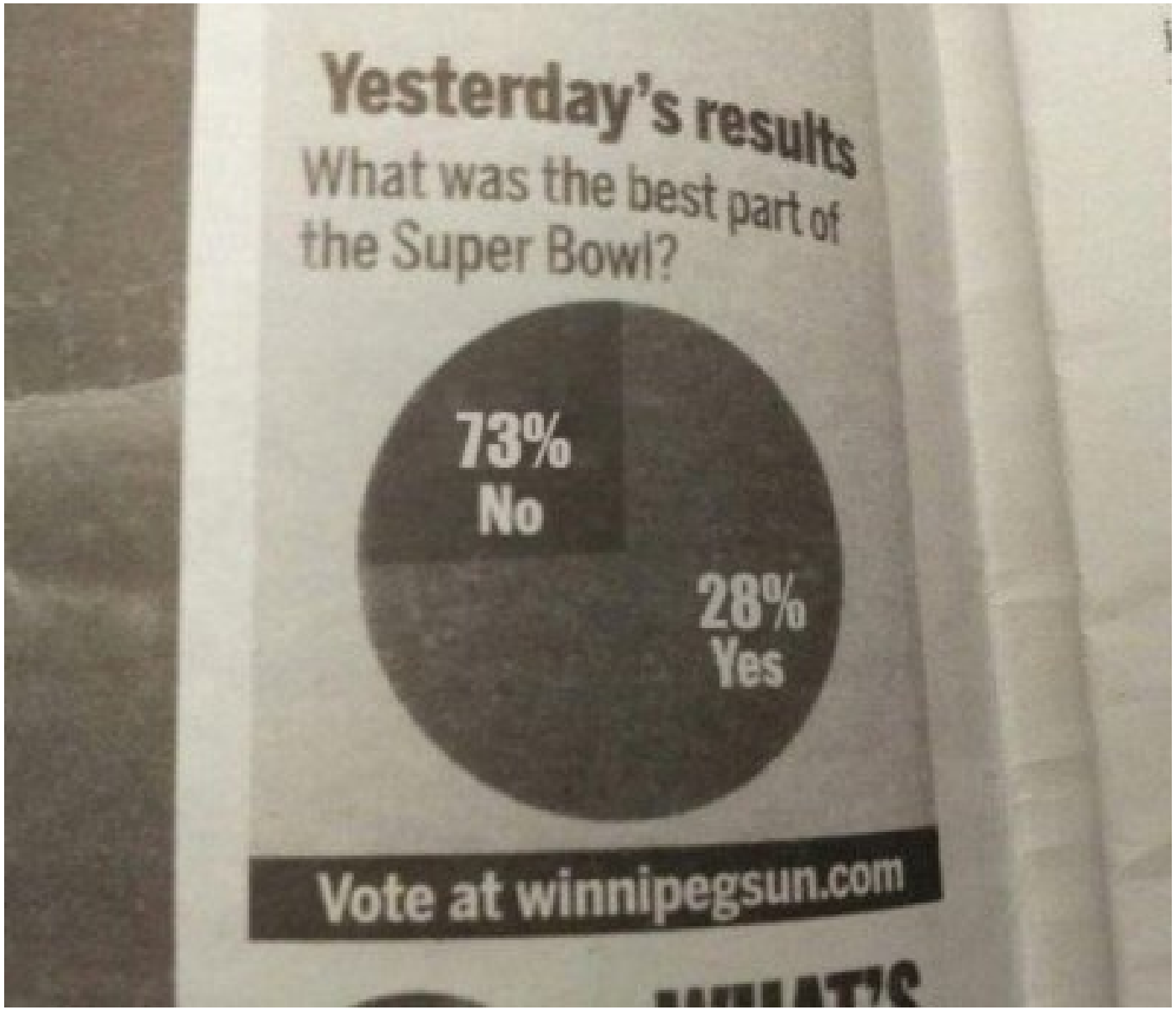
Note that KFC has twice the sales of Starbucks and McDonald's is about 4 times Burger King, but both differences appear much greater.

Misleading graphs: pie chart abuse

Since pie charts represent portions of a whole, the slices should always add up to 100%.



No. Just no.


Yesterday's results
What was the best part of the Super Bowl?
73% No
28% Yes
Vote at winnipegsun.com

Chapter 3
Statistics for Describing, Exploring, and Combining Data

Section 3.2
Measures of Center

Measures of center

In order to understand a data set, values are calculated which summarize the distribution of the data or describe various properties of the data. These are, unsurprisingly, called **descriptive** or **summary** statistics.

Perhaps the most important of these, **measures of center** are a way of representing the value of the middle of the data.

There are four measures of center discussed in this section:

- mean
- median
- mode
- midrange

Mean

The **mean** (the arithmetic mean) is the measure of center calculated by adding the values of the data set and dividing by the size of the data set. Also known as the average.

- Only makes sense with quantitative data
- Sensitive to outliers (extreme or unusual values) and skewed data distributions.

To calculate

Let X be sample of size n of quantitative data with values $x_1, \ldots, x_n$. Then,

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

$\bar{x}$ is the mean of the sample.

- $\sum$ means add all the x_i's, where i is between 1 and n

Mean, example

Example

From the Stat 201 survey, we have the ages of 8 people.

- The sample is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$
- The sample size is $n = 8$
- The sum of the data is

$$\sum x_i = 22 + 32 + 46 + 50 + 33 + 38 + 20 + 24 = 265$$

- The mean is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{265}{8} = 33.125$$

Median

The **median** is the value that is greater than or equal to 50% of the data and less than or equal to 50% of the data.

- Can be used with quantitative and ordinal data
- Not sensitive to extreme values (**resistant** measure of center)

To calculate

Arrange the data in order, from lowest to highest.

- If n is odd, the middle value is the median.
- If n is even, the median is the mean of the two middle values.

Median, example

Example

From the Stat 201 survey, we have the ages of 8 people.

- The sample is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$
- Arranged in order, the sample looks like

$$20 \quad 22 \quad 24 \quad 32 \quad 33 \quad 38 \quad 46 \quad 50$$

- Since n is even, find the the mean of the two middle values.

$$20 \quad 22 \quad 24 \quad \underbrace{32 \quad 33}_{(32+33)/2=32.5} \quad 38 \quad 46 \quad 50$$

- The median is $\tilde{x} = 32.5$

Mean vs. Median

Suppose in our age data set, we replaced the 50 with a 85.

- Mean goes from 33.125 to 37.5
- Median remains unchanged at 32.5

This is why median is called a **resistant** statistic.

- Median is used when we don't want a few extreme values to distort a more reasonable middle, such as house prices or incomes.

Mean vs. Median, cont.

Suppose instead of calculating a grade point average (mean), we calculated a grade point median. Consider a student who got A's in 6 classes and D's in 4.

- The median grade point is 4, an A.
- The GPA for such a student would be 2.8.

The median does not consider all values of a data set. The mean does.

- Mean is used when all values are important or when we expect to have roughly symmetric data.

Mode

The **mode** is the data value with highest frequency.

- Can be used with any kind of data.
- A data set might have more than one mode, or there might not be any mode.

Mode, example

Examples

- The age data, $\{22, 32, 46, 50, 33, 38, 20, 24\}$, has no mode.
- From the survey, favorite kind of taco had these responses:

  {Beef, Beef, Fish, Shrimp, Beef, Pork, Chicken, Beef, Chicken, Beef}

  The mode is "Beef" with a frequency of five.
- Suppose a sample of the class got the following grades on a quiz:

  {A, C, B, A, A, B, C, B}

  The modes are A and B, with frequencies of three each.

Midrange

The **midrange** is the value half way between the minimum and maximum values. Calculate by finding the mean of the minimum and maximum.

- Only makes sense with quantitative data.
- *Very* sensitive the extreme values.
- Easy to calculate, but rarely used.

Example

The age data is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$.

- The minimum age is 20 and the maximum age is 50.
- The midrange is

$$\frac{\min(X) + \max(X)}{2} = \frac{20 + 50}{2} = 35$$

Section 3.3
Measures of Variation

Measures of variation

Another important class of descriptive statistics are **measures of variation** which describe how much the data is spread out.

There are four measures of variation discussed in this section:

- Range
- Variance
- Standard deviation
- Coefficient of variation

Range

The **range** is the difference between the maximum and minimum values.

- Like the midrange, very sensitive to extreme values.

Example

The age data is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$.

- The minimum age is 20 and the maximum age is 50.
- The range is

$$\max(x) - \min(X) = 50 - 20 = 30$$

Variance and standard deviation

The **variance** is the mean of the squared difference of the data from the mean. The **standard deviation** is the square root of the variance.

- More simply, the standard deviation is the average distance of the data from the data mean (the center).
- Always non-negative. A zero standard deviation means all the data are the same value.
- Sensitive to extreme values.
- The units of standard deviation are the same as the data. Variance units are the data units squared.

Variance and standard deviation, calculation

To calculate

Let X be sample of size n of quantitative data with values $x_1, \ldots, x_n$ and sample mean $\bar{x}$. Then,

$$\mathrm{Var}\,(X) = s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \qquad \text{and} \qquad \mathrm{SD}\,(X) = s = \sqrt{s^2}$$

- Note: Never calculate this by hand. Use technology.

Variance and standard deviation, example

Example

The age data is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$. The sample size is $n = 8$ and the sample mean is $\bar{x} = 33.125$

- The variance is

$$\begin{aligned} s^2 &= \frac{\sum (x_i - \bar{x})^2}{n-1} \\ &= \frac{(22 - 33.125)^2 + \cdots + (24 - 33.125)^2}{7} \\ &= 122.125 \end{aligned}$$

- The standard deviation is

$$s = \sqrt{s^2} = \sqrt{122.125} = 11.05$$

Coefficient of variation

The **coefficient of variation** is the standard deviation expressed as a percentege of the mean.

- Useful for comparing samples from two different populations.

To calculate

For sample X with mean $\bar{x}$ and standard deviation s, the coefficient of variation is

$$\frac{s}{\bar{x}} \times 100\%$$

Example

The age data has mean of $\bar{x} = 33.125$ and standard deviation of $s = 11.05$. The coefficient of variation is

$$\frac{11.05}{33.125} \times 100\% = 33.36\%$$

Notation

Remember, values that describe the properties of populations are called **parameters** and values that describe samples are called **statistics**. Notationally, in math formulas or when abbreviating, Greek letters are used to refer to parameters and Latin letters are used to refer to statistics.

Property	Parameter	Statistic
Mean	μ (mu)	$\bar{x}$
Variance	σ^2 (sigma-squared)	s^2
Standard deviation	σ (sigma)	s

Section 3.4
Measures of Relative Standing and Boxplots

Measures of relative standing

Measures of relative standing describe the location of a given data value with a data distribution or a data set.

Two measures of relative standing are discussed in this section:

- Z-scores
- Percentiles

Z-scores

A **z-score** describes the relative position of a data value within a data distribution.

- Another way to put it is a z-score is the number of standard deviations that a particular value is above or below the mean.
- Z-scores are standardized, so they can be used to compare values from different populations.
- A positive z-score means the value is greater than the mean and a negative z-score means that it is below the mean.
- Z-scores can be calculated for samples or populations, if the population mean and standard deviation are known.

Z-scores, calculations

To calculate

- For a sample X with sample mean $\bar{x}$ and standard deviation s, the z-score for a value x is

$$z = \frac{x - \bar{x}}{s}$$

- For a population $\mathcal{P}$ with population mean μ and standard deviation σ, the z-score for value x is

$$z = \frac{x - \mu}{\sigma}$$

Z-scores, example

Example

The age data has mean of $\bar{x} = 33.125$ and SD of $s = 11.05$.

- Suppose a new student joins the class. His age is 44. He has an age z-score of

$$z = \frac{x - \bar{x}}{s} = \frac{44 - 33.125}{11.05} = \frac{10.875}{11.05} = 0.984$$

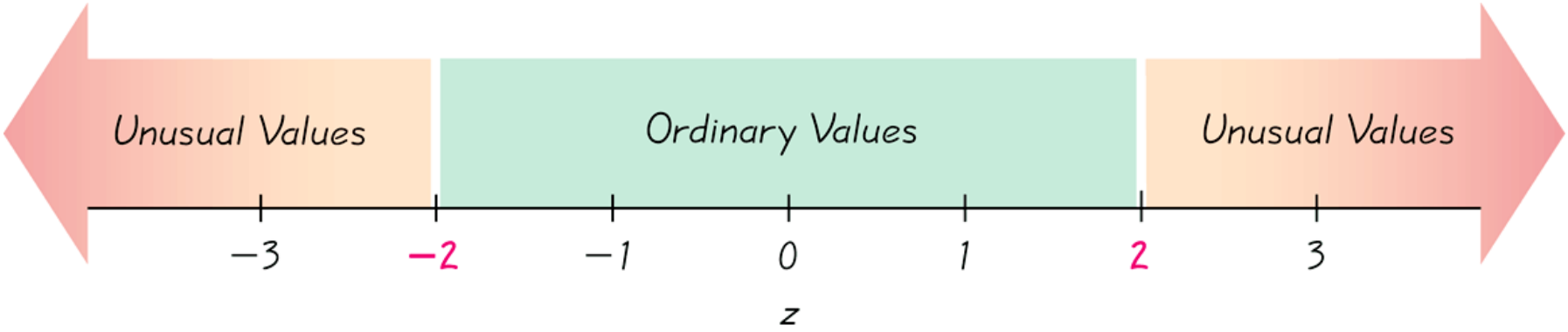
  His age is almost a standard deviation above the class mean.

- Another student joins the class. Her age is 27. She has an age z-score of

$$z = \frac{x - \bar{x}}{s} = \frac{27 - 33.125}{11.05} = \frac{-6.125}{11.05} = -0.554$$

  Her age is about a half standard deviation below the class mean.

Unusual values



A value is called **unusual** if it has a z-score z such that $z < -2$ or $z > 2$.
A value is **ordinary** if z is between -2 and 2.

Example

Our two new students with z-scores of 0.984 and -0.554 both have ordinary ages. A third new student aged 85, with a z-score of $z = (85 - 33.125)/11.05 = 4.69$, has an unusual age.

Percentiles

Percentiles measure relative position within a data set as order rank. In other words, the value at the pth percentile (written as P_p) in a data set is greater than $p\%$ of the data.

To calculate

To find the percentile of a value x in a data set,

$$\%\text{ile} = \frac{\text{number of values} < x}{n} \times 100\%$$

To find the value of P_p (the pth percentile), calculate the rank,

$$r = \frac{p}{100} \times n$$

If r is a whole number, P_p is the mean of the rth and $(r+1)$th values. If not, round up. Then, P_p is the rth value in an ordered list.

Percentile, example

Example

The age data, in order is,

$$20 \quad 22 \quad 24 \quad 32 \quad 33 \quad 38 \quad 46 \quad 50$$

- The percentile of the value 38 is

$$\frac{\text{number of values} < x}{n} \times 100\% = \frac{5}{8} \times 100\% = 62.5\,\% \Rightarrow P_{63}$$

- To find the 30th percentile, P_{30}, calculate rank

$$r = \frac{p}{100} \times n = \frac{30}{100} \times 8 = 2.4$$

Round up r to 3. P_{30} is the 3rd value, 24.

Quartiles

The **quartiles** are values that divide the data set into 4 parts, or quarters.

$$Q_1 = P_{25} \qquad Q_2 = P_{50} \qquad Q_3 = P_{75}$$

- Note: The median is equivalent to Q_2 and P_{50}.

5 number summary

The **5 number summary** summarize the distribution of a data set.

The 5 numbers are:

- Minimum
- Q_1
- Median (or Q_2)
- Q_3
- Maximum

5 number summary, example

Example

The age data, in order is,

$$20 \quad 22 \quad 24 \quad 32 \quad 33 \quad 38 \quad 46 \quad 50$$

The 5 number summary is

$$\underbrace{20}_{\min} \quad \underbrace{23}_{Q_1} \quad \underbrace{32.5}_{\text{med}} \quad \underbrace{42}_{Q_3} \quad \underbrace{50}_{\max}$$

Boxplots

A **boxplot** is a graph depicting the 5 number summary.