Chapter 1

Introduction to Statistics

What is Data?

Data

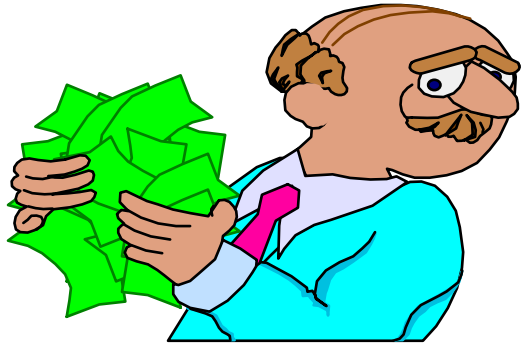
Consist of information coming from observations, counts, measurements, or responses.

- "People who eat three daily servings of whole grains have been shown to reduce their risk of…stroke by 37%." *(Source: Whole Grains Council)*
- "Seventy percent of the 1500 U.S. spinal cord injuries to minors result from vehicle accidents, and 68 percent were not wearing a seatbelt." *(Source: UPI)*

What is Statistics?

Statistics

The science of collecting, organizing, analyzing, and interpreting data in order to make decisions.



Descriptive Statistics: *Who was the best baseball player of all time?*

Correlation: *How does Netflix know what movies I like?*

Basic Probability: *Don't buy the extended warranty.*

The Central Limit Theorem: *The Tiger Woods of statistics*

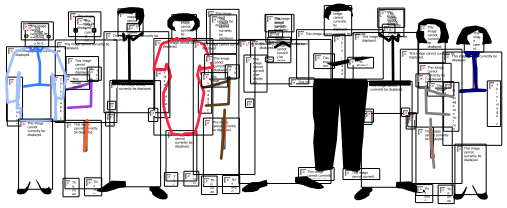
Inference: *Why my statistics professor thought I might have cheated*

Polling/sampling: *How we know that 64 percent of Americans support the death penalty*

Data Sets

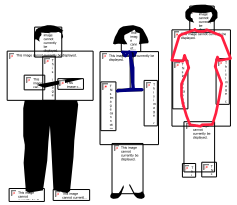
Population

The collection of ***all*** outcomes, responses, measurements, or counts that are of interest.



Sample

A subset, or part, of the population.

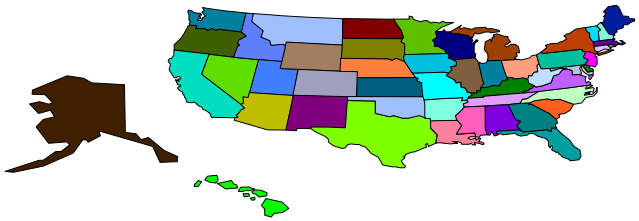


Parameter and Statistic

Parameter

A numerical description of a population characteristic.

Average age of all people in the United States



Statistic

A numerical description of a sample characteristic.

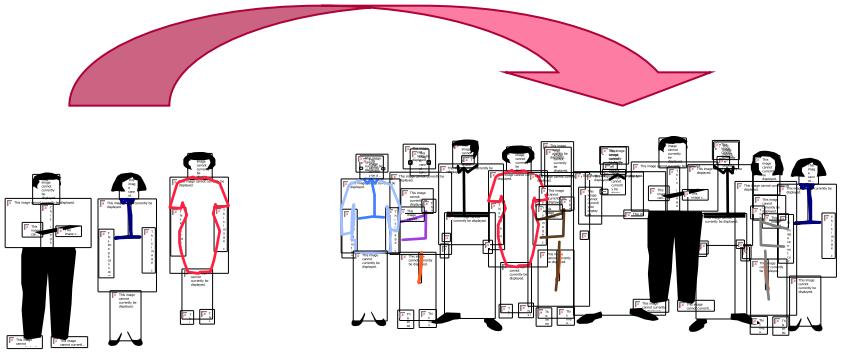
Average age of people from a sample of three states

Branches of Statistics

Descriptive Statistics
Involves organizing, summarizing, and displaying data.

e.g. Tables, charts, averages

Inferential Statistics
Involves using ***sample data*** to draw conclusions about a ***population.***

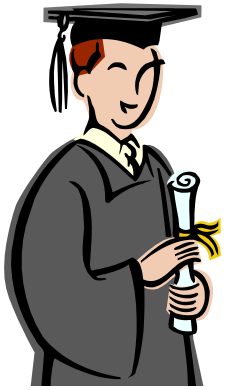


Types of Data

Qualitative Data

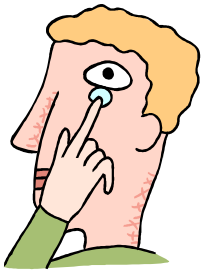
Consists of attributes, labels, or nonnumerical entries.

Major



Place of birth

Eye color



Types of Data

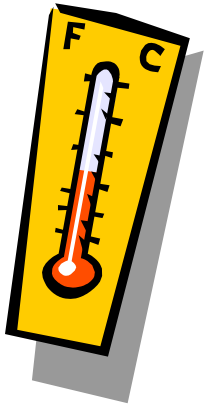
Quantitative data

Numerical measurements or counts.

Age

Weight of a letter

Temperature



Levels of Measurement

Nominal level of measurement

- Qualitative data only
- Categorized using names, labels, or qualities
- No mathematical computations can be made

Ordinal level of measurement

- Qualitative or quantitative data
- Data can be arranged in order, or ranked
- Differences between data entries is not meaningful

Levels of Measurement

Interval level of measurement

- Quantitative data
- Data can ordered
- Differences between data entries is meaningful
- Zero represents a position on a scale (not an inherent zero – zero does not imply “none”)

Levels of Measurement

Ratio level of measurement

- Similar to interval level
- Zero entry is an inherent zero (implies “none”)
- A ratio of two data values can be formed
- One data value can be expressed as a multiple of another

Summary of Four Levels of Measurement

Level of Measurement	Put data in categories	Arrange data in order	Subtract data values	Determine if one data value is a multiple of another
Nominal	Yes	No	No	No
Ordinal	Yes	Yes	No	No
Interval	Yes	Yes	Yes	No
Ratio	Yes	Yes	Yes	Yes

Section 1.3 Objectives

- How to design a statistical study and how to distinguish between an observational study and an experiment
- How to collect data by using a survey or a simulation
- How to design an experiment
- How to create a sample using random sampling, simple random sampling, stratified sampling, cluster sampling, and systematic sampling and how to identify a biased sample

Designing a Statistical Study

1. Identify the variable(s) of interest (the focus) and the population of the study.
2. Develop a detailed plan for collecting data. If you use a sample, make sure the sample is representative of the population.
3. Collect the data.
4. Describe the data using descriptive statistics techniques.
5. Interpret the data and make decisions about the population using inferential statistics.
6. Identify any possible errors.

Data Collection

Observational study

- A researcher observes and measures characteristics of interest of part of a population.

- Researchers observed and recorded the mouthing behavior on nonfood objects of children up to three years old. *(Source: Pediatric Magazine)*

Data Collection

Experiment

- A treatment is applied to part of a population and responses are observed.

- An experiment was performed in which diabetics took cinnamon extract daily while a control group took none. After 40 days, the diabetics who had the cinnamon reduced their risk of heart disease while the control group experienced no change. *(Source: Diabetes Care)*

Data Collection

Simulation

- Uses a mathematical or physical model to reproduce the conditions of a situation or process.
- Often involves the use of computers.

- Automobile manufacturers use simulations with dummies to study the effects of crashes on humans.

Data Collection

Survey

- An investigation of one or more characteristics of a population.
- Commonly done by interview, Internet, phone, or mail.

- A survey is conducted on a sample of female physicians to determine whether the primary reason for their career choice is financial stability.

Key Elements of Experimental Design

- Control
- Randomization
- Sample Size
- Replication

Key Elements of Experimental Design: Sample Size

- **Sample Size**
  - The number of subjects in a study is very important to experimental design.

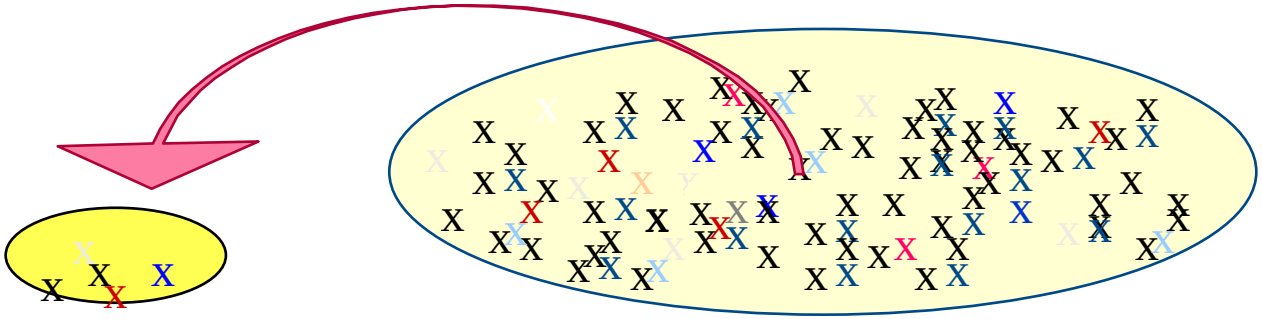
Sampling Techniques

Random Sample

Every member of the population has an equal chance of being selected.

Simple Random Sample

Every possible sample of the same size has the same chance of being selected.



Simple Random Sample

- Random numbers can be generated by a random number table, a software program or a calculator.
- Assign a number to each member of the population.
- Members of the population that correspond to these numbers become members of the sample.

Other Sampling Techniques

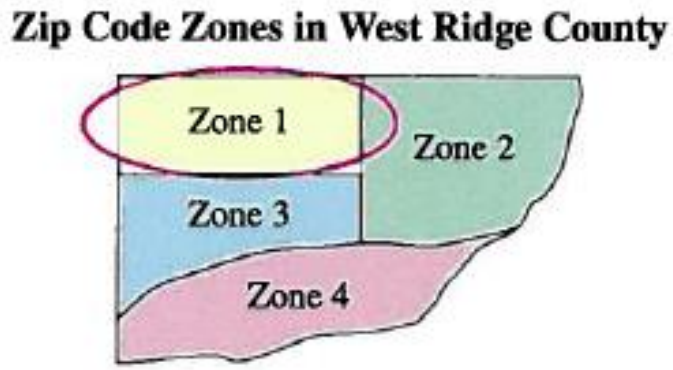
Stratified Sample

- Divide a population into groups (strata) and select a random sample from each group.
- To collect a stratified sample of the number of people who live in West Ridge County households, you could divide the households into socioeconomic levels and then randomly select households from each level.

Other Sampling Techniques

Cluster Sample

- Divide the population into groups (clusters) and select all of the members in one or more, but not all, of the clusters.
- In the West Ridge County example you could divide the households into clusters according to zip codes, then select all the households in one or more, but not all, zip codes.

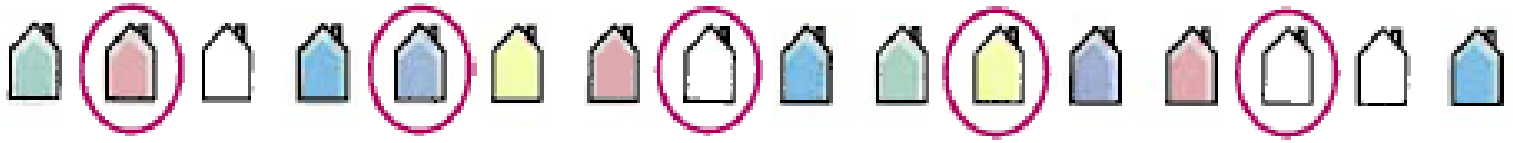


Other Sampling Techniques

Systematic Sample

- Choose a starting value at random. Then choose every k^{th} member of the population.

- In the West Ridge County example you could assign a different number to each household, randomly choose a starting number, then select every 100^{th} household.



Other Sampling Techniques

Convenience Sample

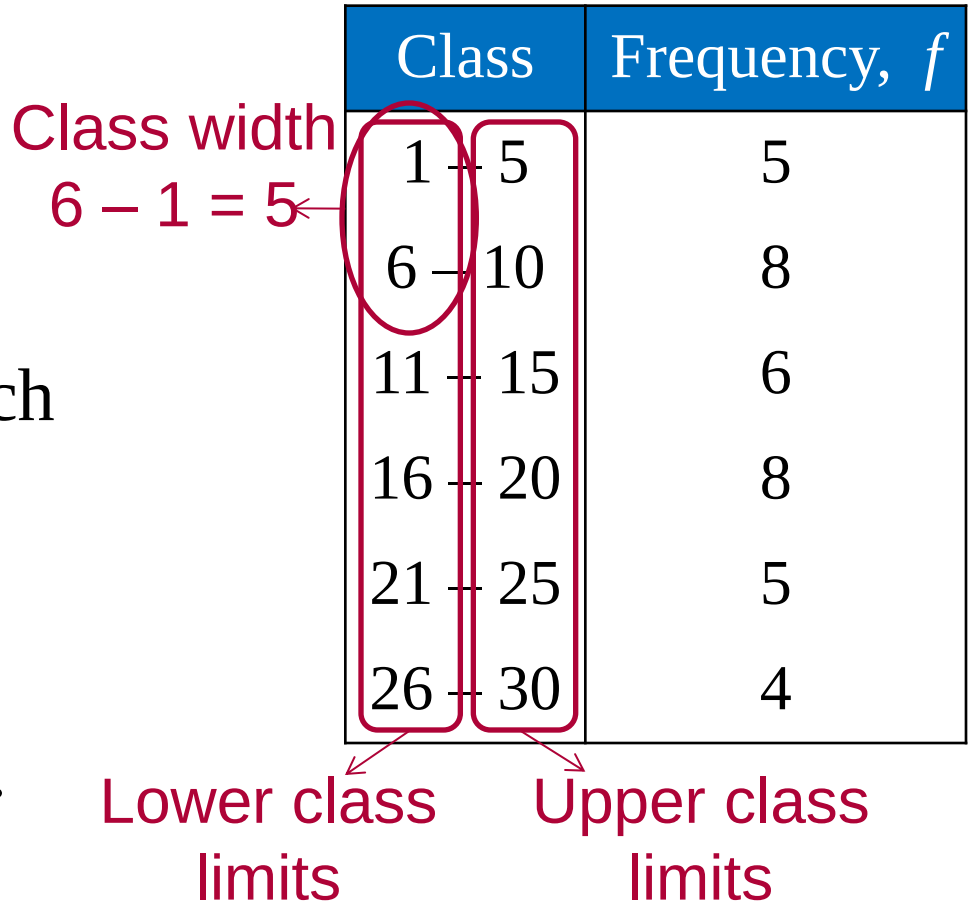
- Choose only members of the population that are easy to get
- Often leads to biased studies (not recommended)

Frequency Distributions and Their Graphs

Frequency Distribution

Frequency Distribution

- A table that shows **classes** or **intervals** of data with a count of the number of entries in each class.
- The **frequency, *f*,** of a class is the number of data entries in the class.



Constructing a Frequency Distribution

1. Decide on the number of classes.
   - Usually between 5 and 20; otherwise, it may be difficult to detect any patterns.
2. Find the class width.
   - Determine the range of the data.
   - Divide the range by the number of classes.
   - *Round up to the next convenient number.*

Constructing a Frequency Distribution

3. Find the class limits.
   - You can use the minimum data entry as the lower limit of the first class.
   - Find the remaining lower limits (add the class width to the lower limit of the preceding class).
   - Find the upper limit of the first class. Remember that classes cannot overlap.
   - Find the remaining upper class limits.

Constructing a Frequency Distribution

4. Make a tally mark for each data entry in the row of the appropriate class.
5. Count the tally marks to find the total frequency *f* for each class.

Determining the Cumulative Frequency

Cumulative frequency of a class

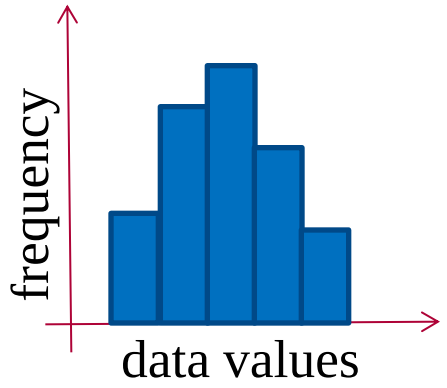
- The sum of the frequency for that class and all previous classes.

Class	Frequency, *f*	Cumulative frequency
59 – 114	5	5
115 – 170	+ 8	13
171 – 226	+ 6	19

Graphs of Frequency Distributions

Frequency Histogram

- A bar graph that represents the frequency distribution.
- The horizontal scale is quantitative and measures the data values.
- The vertical scale measures the frequencies of the classes.
- Consecutive bars must touch.

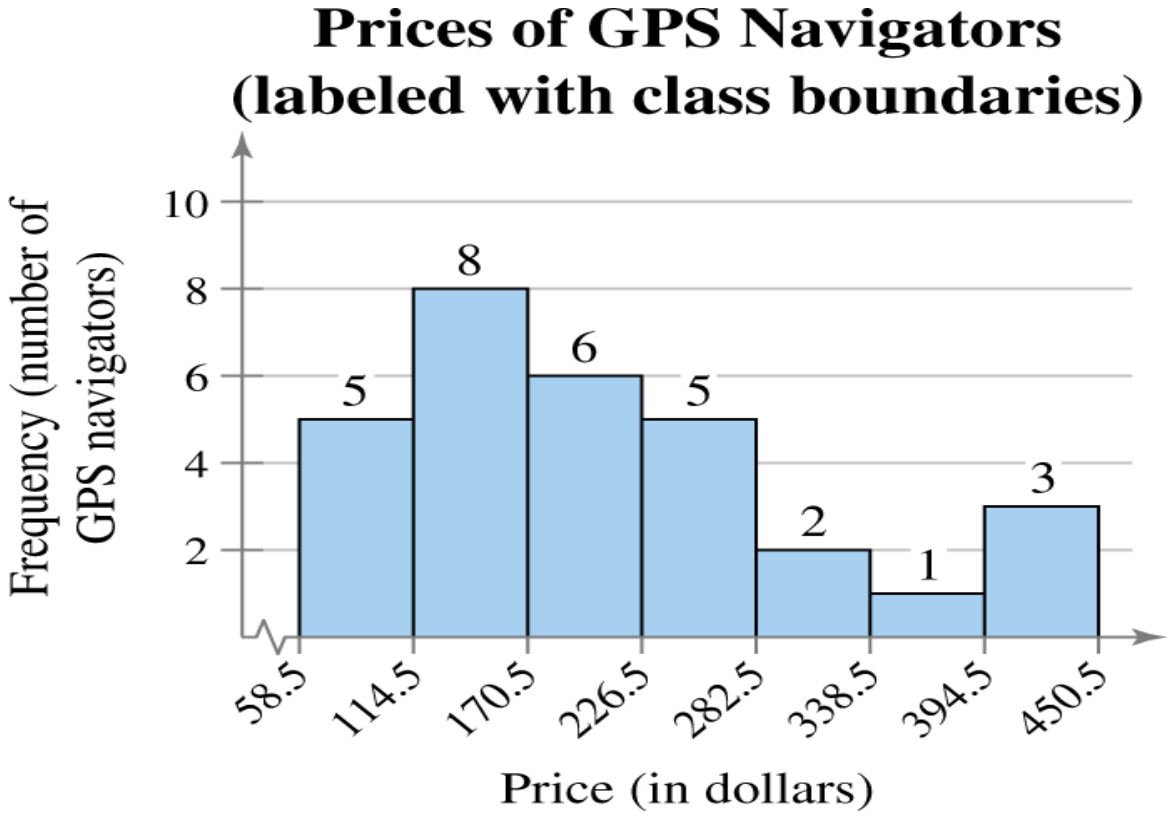


Example: Frequency Histogram

Construct a frequency histogram for the global positioning system (GPS) navigators.

Class	Class boundaries	Midpoint	Frequency, f
59 – 114	58.5 – 114.5	86.5	5
115 – 170	114.5 – 170.5	142.5	8
171 – 226	170.5 – 226.5	198.5	6
227 – 282	226.5 – 282.5	254.5	5
283 – 338	282.5 – 338.5	310.5	2
339 – 394	338.5 – 394.5	366.5	1
395 – 450	394.5 – 450.5	422.5	3

Solution: Frequency Histogram (using class boundaries)



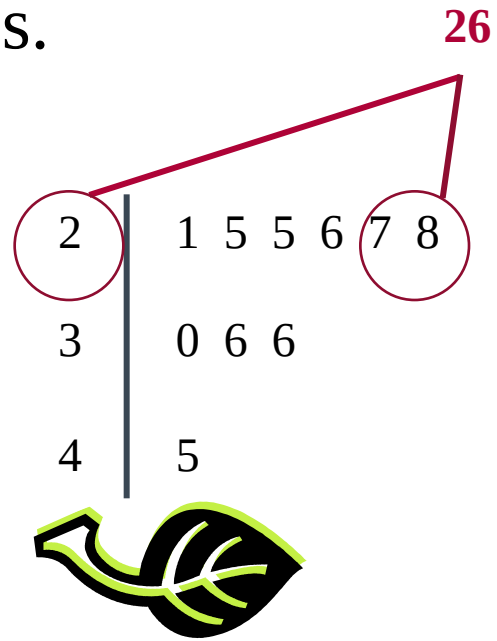
You can see that more than half of the GPS navigators are priced below $226.50.

Graphing Quantitative Data Sets

Stem-and-leaf plot

- Each number is separated into a **stem** and a **leaf**.
- Similar to a histogram.
- Still contains original data values.

Data: 21, 25, 25, **26**, 27, 28,
30, 36, 36, 45

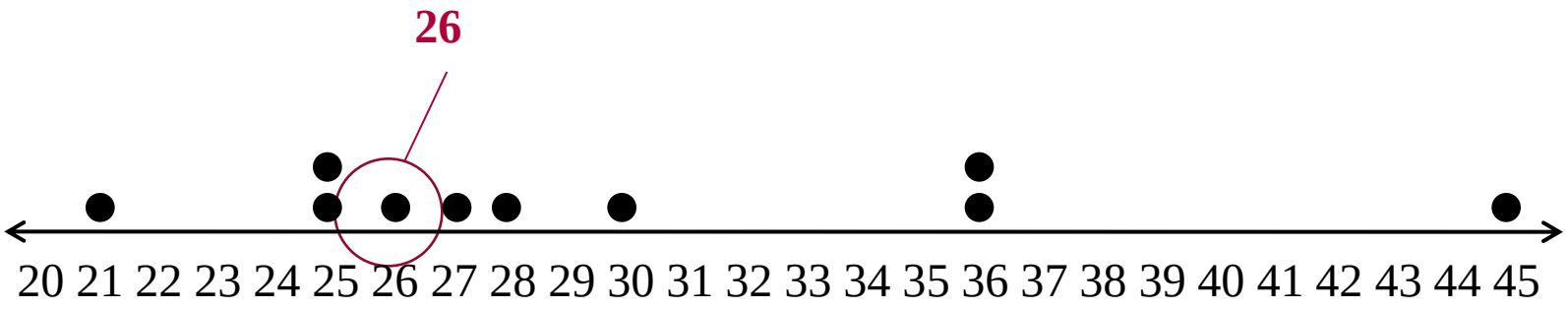


Graphing Quantitative Data Sets

Dot plot

- Each data entry is plotted, using a point, above a horizontal axis

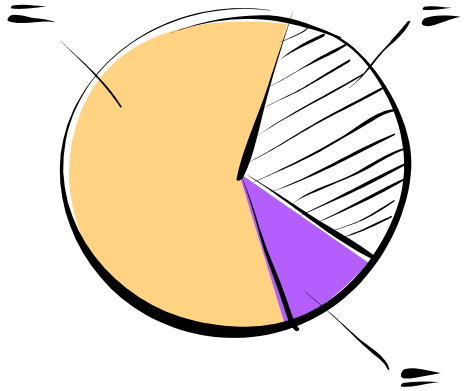
Data: 21, 25, 25, **26**, 27, 28, 30, 36, 36, 45



Graphing Qualitative Data Sets

Pie Chart

- A circle is divided into sectors that represent categories.
- The area of each sector is proportional to the frequency of each category.



Example: Constructing a Pie Chart

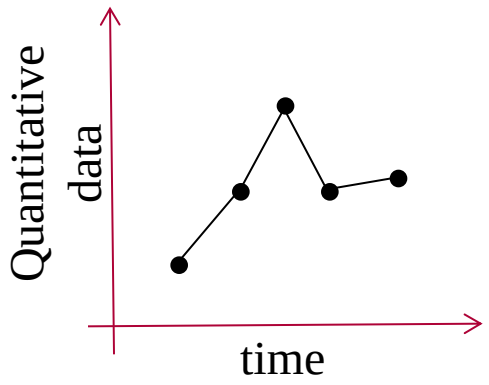
The numbers of earned degrees conferred (in thousands) in 2007 are shown in the table. Use a pie chart to organize the data. ***(Source: U.S. National Center for Educational Statistics)***

Type of degree	Number (thousands)
Associate's	728
Bachelor's	1525
Master's	604
First professional	90
Doctoral	60

Graphing Paired Data Sets

Time Series

- Data set is composed of quantitative entries taken at regular intervals over a period of time.
  - e.g., The amount of precipitation measured each day for one month.
- Use a **time series chart** to graph.



Chapter 2
Descriptive Statistics

Measures of Central Tendency

Measures of Central Tendency

Measure of central tendency

- A value that represents a typical, or central, entry of a data set.
- Most common measures of central tendency:
  - Mean
  - Median
  - Mode

Measure of Central Tendency: Mean

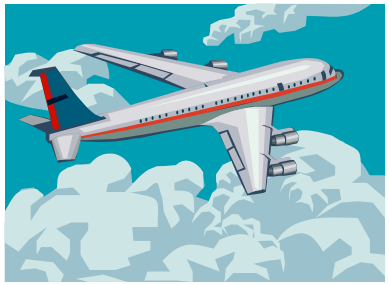
Mean (average)

- The sum of all the data entries divided by the number of entries.
- **Sigma notation**: Σx = add all of the data entries (x) in the data set.
- **Population mean**: $\mu = \frac{\Sigma x}{N}$
- **Sample mean**: $\bar{x} = \frac{\Sigma x}{n}$

Example: Finding a Sample Mean

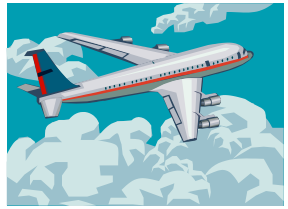
The prices (in dollars) for a sample of roundtrip flights from Chicago, Illinois to Cancun, Mexico are listed. What is the mean price of the flights?

872 432 397 427 388 782 397



Solution: Finding a Sample Mean

872 432 397 427 388 782 397



- The sum of the flight prices is

$\Sigma x = 872 + 432 + 397 + 427 + 388 + 782 + 397 = 3695$

- To find the mean price, divide the sum of the prices by the number of prices in the sample

$$\bar{x} = \frac{\Sigma x}{n} = \frac{3695}{7} \approx 527.9$$

The mean price of the flights is about $527.90.

Measure of Central Tendency: Median

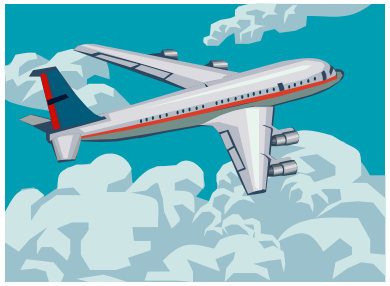
Median

- The value that lies in the middle of the data when the data set is **ordered**.
- Measures the center of an ordered data set by dividing it into two equal parts.
- If the data set has an
  - **odd number of entries**: median is the middle data entry.
  - **even number of entries**: median is the mean of the two middle data entries.

Example: Finding the Median

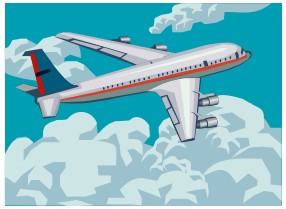
The prices (in dollars) for a sample of roundtrip flights from Chicago, Illinois to Cancun, Mexico are listed. Find the median of the flight prices.

872 432 397 427 388 782 397



Solution: Finding the Median

872 432 397 427 388 782 397



- First order the data.

  388 397 397 427 432 782 872

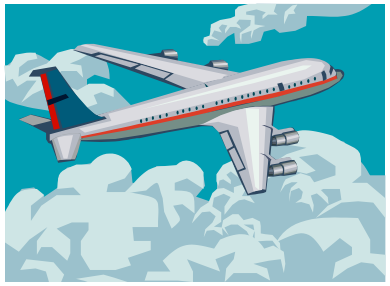
- There are seven entries (an odd number), the median is the middle, or fourth, data entry.

The median price of the flights is $427.

Example: Finding the Median

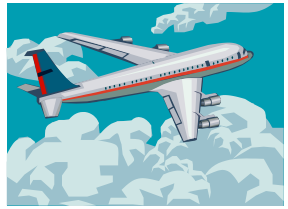
The flight priced at $432 is no longer available. What is the median price of the remaining flights?

872 397 427 388 782 397



Solution: Finding the Median

872 397 427 388 782 397



- First order the data.

  388 397 397 427 782 872

- There are six entries (an even number), the median is the mean of the two middle entries.

$$\text{Median} = \frac{397 + 427}{2} = 412$$

The median price of the flights is $412.

Measure of Central Tendency: Mode

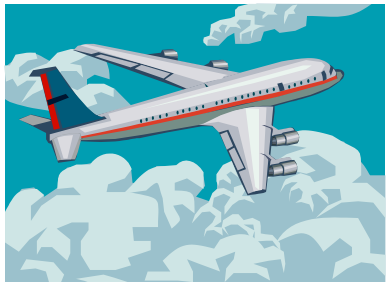
Mode

- The data entry that occurs with the greatest frequency.
- If no entry is repeated the data set has no mode.
- If two entries occur with the same greatest frequency, each entry is a mode (**bimodal**).

Example: Finding the Mode

The prices (in dollars) for a sample of roundtrip flights from Chicago, Illinois to Cancun, Mexico are listed. Find the mode of the flight prices.

872 432 397 427 388 782 397



Comparing the Mean, Median, and Mode

- All three measures describe a typical entry of a data set.
- Advantage of using the mean:
  - The mean is a reliable measure because it takes into account every entry of a data set.
- Disadvantage of using the mean:
  - Greatly affected by **outliers** (a data entry that is far removed from the other entries in the data set).

Example: Comparing the Mean, Median, and Mode

Find the mean, median, and mode of the sample ages of a class shown. Which measure of central tendency best describes a typical entry of this data set? Are there any outliers?

Ages in a class						
20	20	20	20	20	20	21
21	21	21	22	22	22	23
23	23	23	24	24	65	

Solution: Comparing the Mean, Median, and Mode

Ages in a class						
20	20	20	20	20	20	21
21	21	21	22	22	22	23
23	23	23	24	24	65	

Mean: $\bar{x} = \frac{\Sigma x}{n} = \frac{20 + 20 + \ldots + 24 + 65}{20} \approx 23.8$ years

Median: $\frac{21 + 22}{2} = 21.5$ years

Mode: 20 years (the entry occurring with the greatest frequency)

.

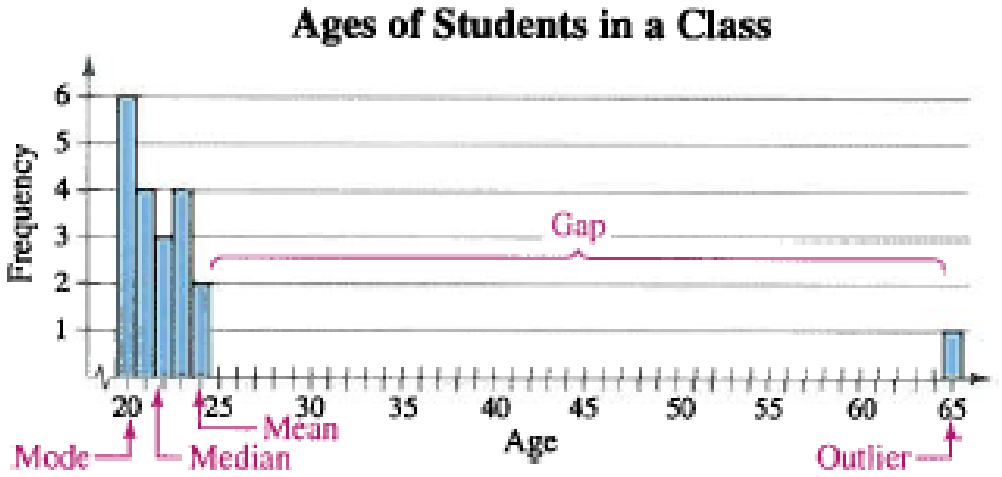
Solution: Comparing the Mean, Median, and Mode

Mean ≈ 23.8 years Median = 21.5 years Mode = 20 years

- The mean takes every entry into account, but is influenced by the **outlier** of 65.
- The median also takes every entry into account, and it is not affected by the outlier.
- In this case the mode exists, but it doesn't appear to represent a typical entry.

Solution: Comparing the Mean, Median, and Mode

Sometimes a graphical comparison can help you decide which measure of central tendency best represents a data set.



In this case, it appears that the **median** best describes the data set.

.

Mean of Grouped Data

Mean of a Frequency Distribution

- Approximated by

$$\bar{x} = \frac{\Sigma(x \cdot f)}{n} \qquad n = \Sigma f$$

where x and f are the midpoints and frequencies of a class, respectively

Finding the Mean of a Frequency Distribution

In Words	*In Symbols*
1. Find the midpoint of each class.	$x = \frac{\text{(lower limit)+(upper limit)}}{2}$
2. Find the sum of the products of the midpoints and the frequencies.	$\Sigma(x \cdot f)$
3. Find the sum of the frequencies.	$n = \Sigma f$
4. Find the mean of the frequency distribution.	$\bar{x} = \frac{\Sigma(x \cdot f)}{n}$

Example: Find the Mean of a Frequency Distribution

Use the frequency distribution to approximate the mean number of minutes that a sample of Internet subscribers spent online during their most recent session.

Class	Midpoint	Frequency, f
7 – 18	12.5	6
19 – 30	24.5	10
31 – 42	36.5	13
43 – 54	48.5	8
55 – 66	60.5	5
67 – 78	72.5	6
79 – 90	84.5	2

Solution: Find the Mean of a Frequency Distribution

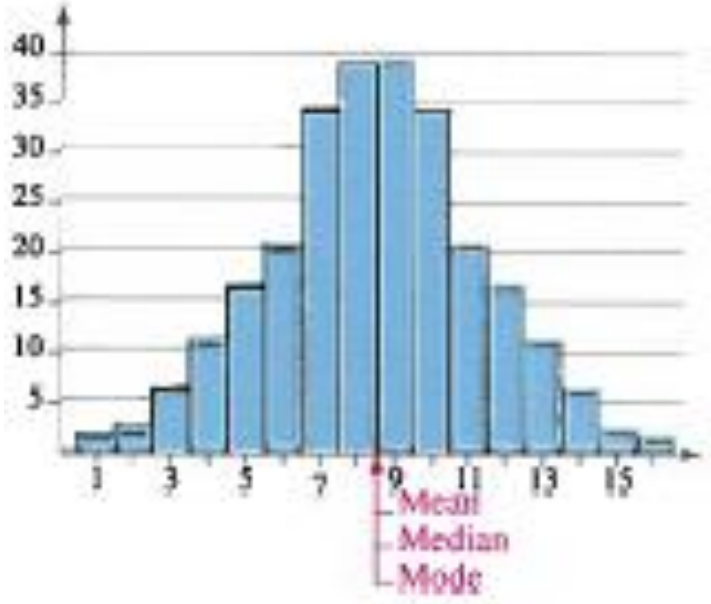
Class	Midpoint, x	Frequency, f	$(x \cdot f)$
7 – 18	12.5	6	$12.5 \cdot 6 = 75.0$
19 – 30	24.5	10	$24.5 \cdot 10 = 245.0$
31 – 42	36.5	13	$36.5 \cdot 13 = 474.5$
43 – 54	48.5	8	$48.5 \cdot 8 = 388.0$
55 – 66	60.5	5	$60.5 \cdot 5 = 302.5$
67 – 78	72.5	6	$72.5 \cdot 6 = 435.0$
79 – 90	84.5	2	$84.5 \cdot 2 = 169.0$
		n = 50	$\Sigma(x \cdot f) = 2089.0$

$$\bar{x} = \frac{\Sigma(x \cdot f)}{n} = \frac{2089}{50} \approx 41.8 \text{ minutes}$$

The Shape of Distributions

Symmetric Distribution

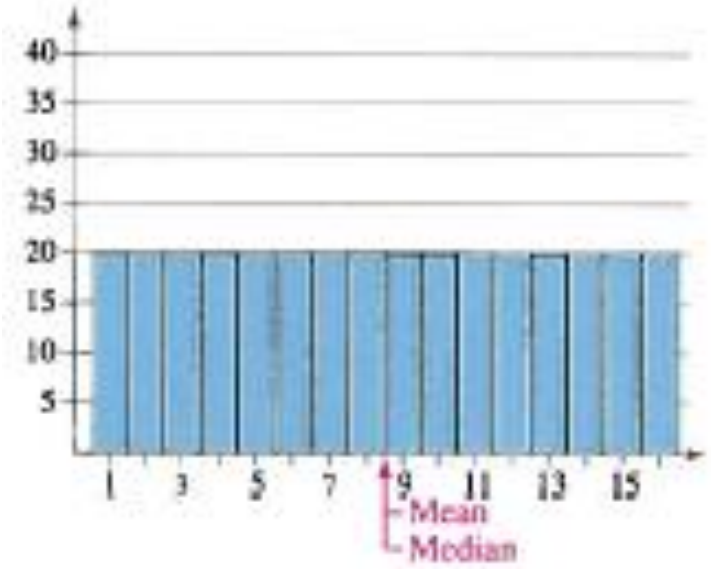
- A vertical line can be drawn through the middle of a graph of the distribution and the resulting halves are approximately mirror images.



The Shape of Distributions

Uniform Distribution (rectangular)

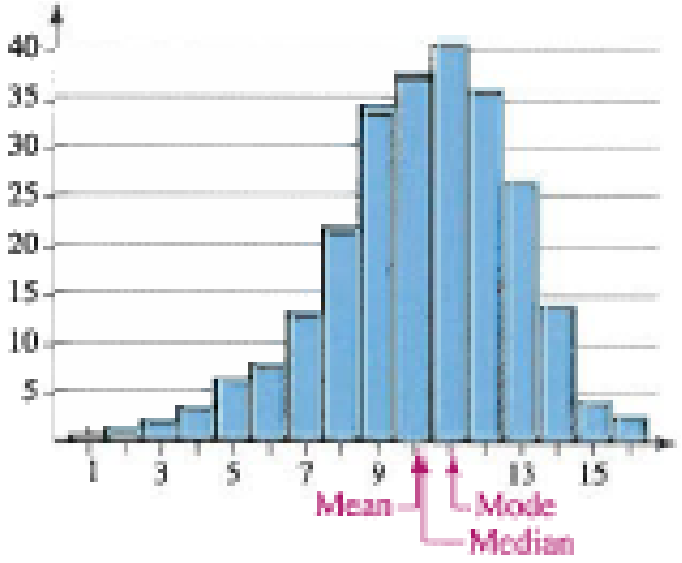
- All entries or classes in the distribution have equal or approximately equal frequencies.
- Symmetric.



The Shape of Distributions

Skewed Left Distribution (negatively skewed)

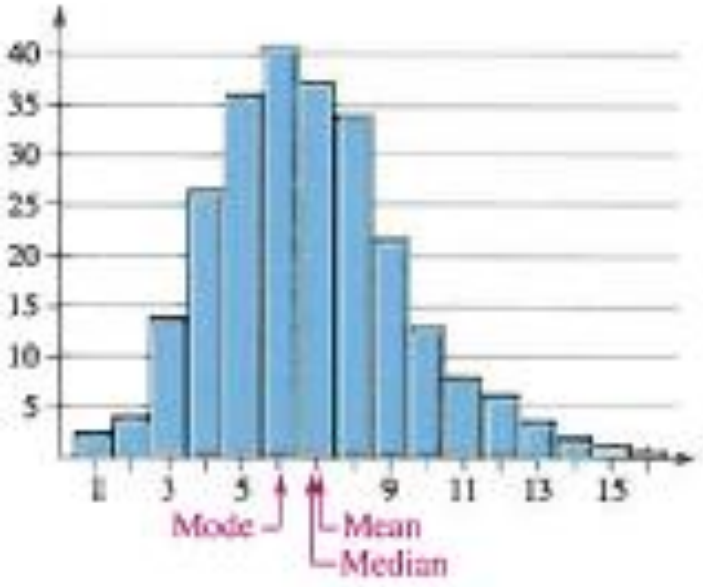
- The “tail” of the graph elongates more to the left.
- The mean is to the left of the median.



The Shape of Distributions

Skewed Right Distribution (positively skewed)

- The “tail” of the graph elongates more to the right.
- The mean is to the right of the median.



Measures of Variation

Because this section introduces the concept of variation, this is one of the most important sections in the entire book

Definition

The range of a set of data is the difference between the highest value and the lowest value

highest value **–** **lowest value**

Definition

The standard deviation of a set of sample values is a measure of variation of values about the mean

Sample Standard Deviation Formula

$$S = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n - 1}}$$

Standard Deviation - Key Points

- The standard deviation is a measure of variation of all values from the mean
- The value of the standard deviation s is usually positive
- The value of the standard deviation s can increase dramatically with the inclusion of one or more outliers (data values far away from all others)
- The units of the standard deviation s are the same as the units of the original data values

Population Standard Deviation

$$\sigma = \sqrt{\frac{\Sigma (x - \mu)^2}{N}}$$

This formula is similar to Formula 2-4, but instead the population mean and population size are used

Definition

- The **variance** of a set of values is a measure of variation equal to the square of the standard deviation.

- Sample variance: Square of the sample standard deviation **s**

- Population variance: Square of the population standard deviation σ

Estimation of Standard Deviation
Range Rule of Thumb

For interpreting a known value of the standard deviation s, find rough estimates of the minimum and maximum "usual" values by using:

Minimum "usual" value $\approx$ (mean) – 2 X (standard deviation)

Maximum "usual" value $\approx$ (mean) + 2 X (standard deviation)

Definition

Empirical (68-95-99.7) Rule

For data sets having a distribution that is approximately bell shaped, the following properties apply:

- **About 68% of all values fall within 1 standard deviation of the mean**
- **About 95% of all values fall within 2 standard deviations of the mean**
- **About 99.7% of all values fall within 3 standard deviations of the mean**

The Empirical Rule

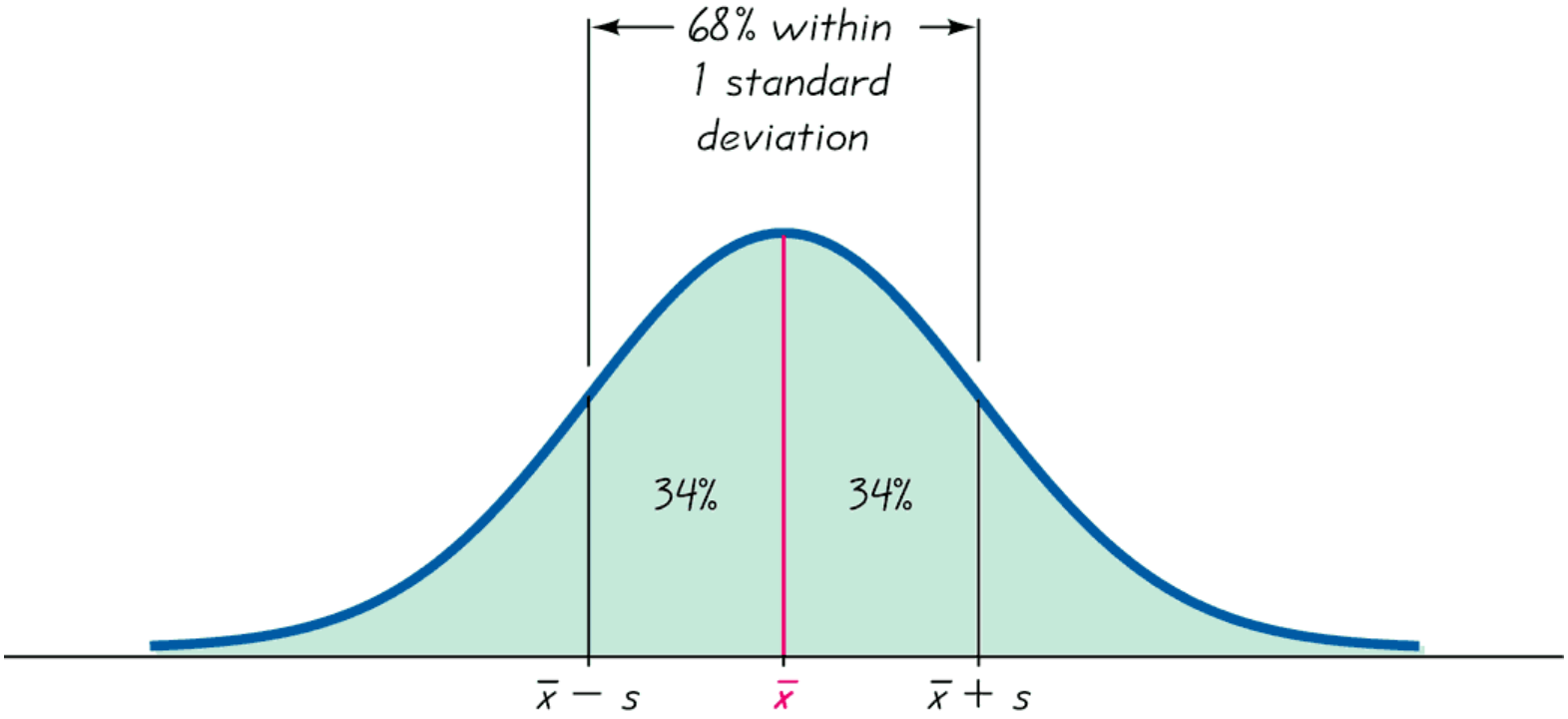


FIGURE 2-13

The Empirical Rule

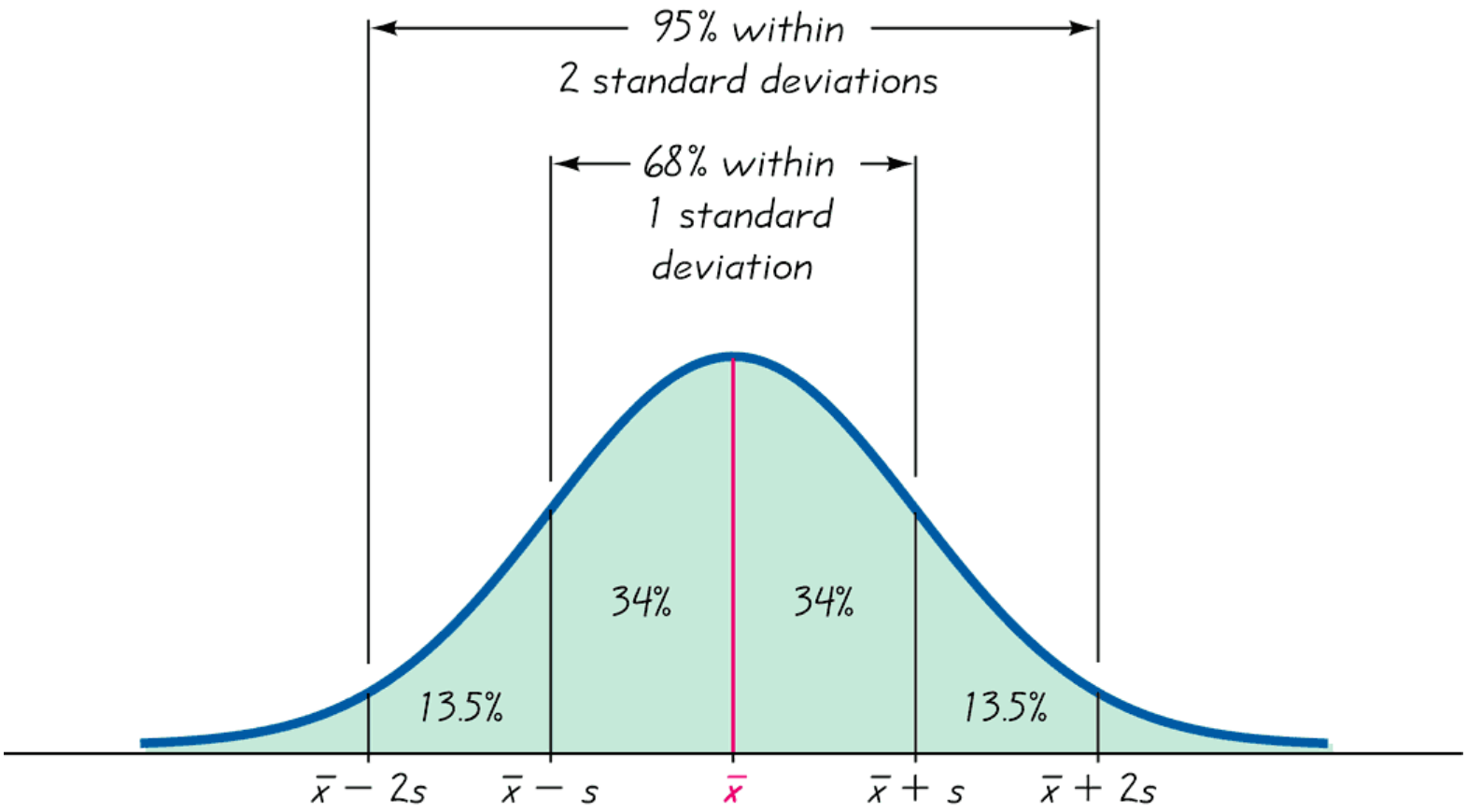


FIGURE 2-13

The Empirical Rule

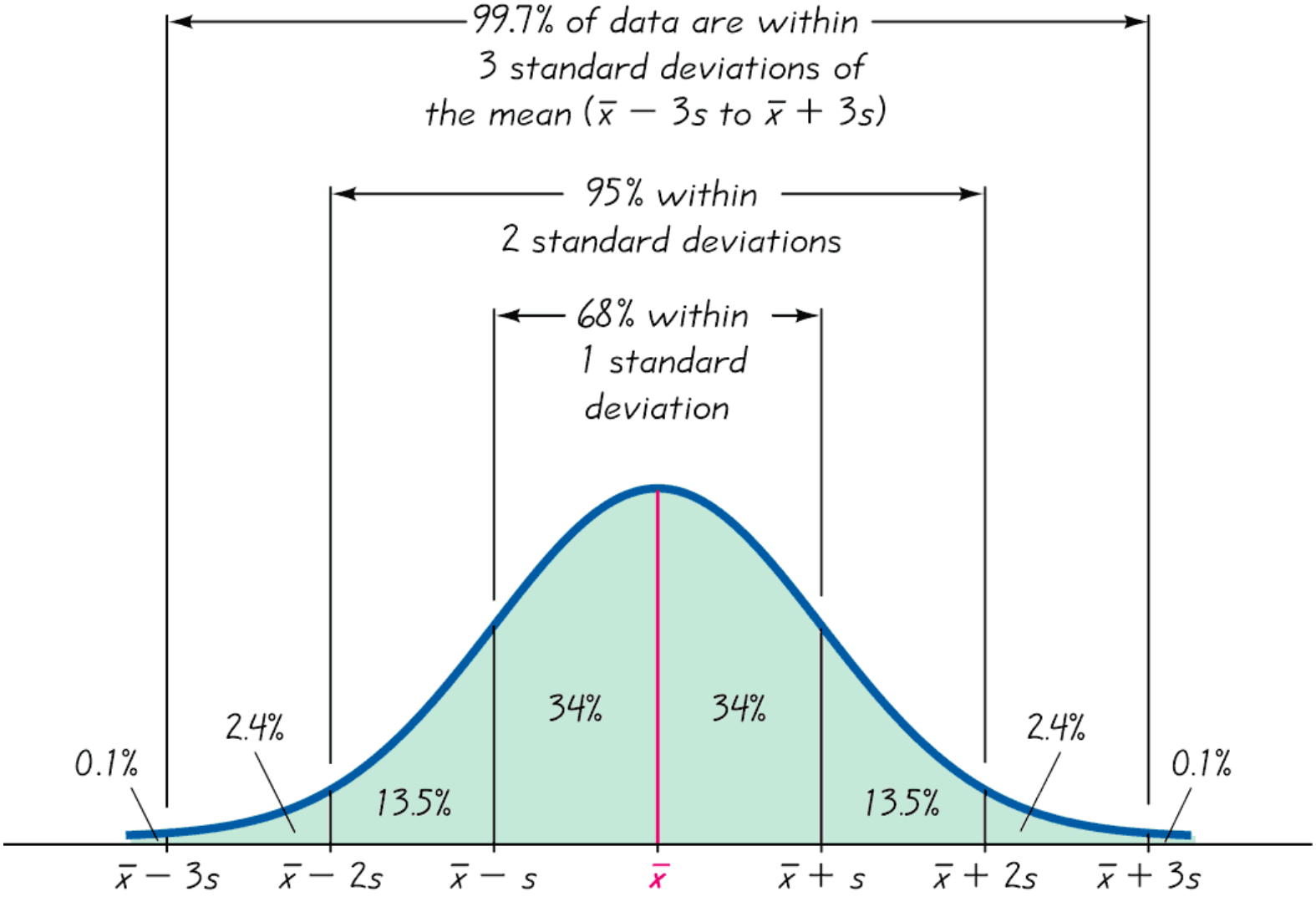


FIGURE 2-13