
WEEK 1 - LECTURE 1

Summarizing Data

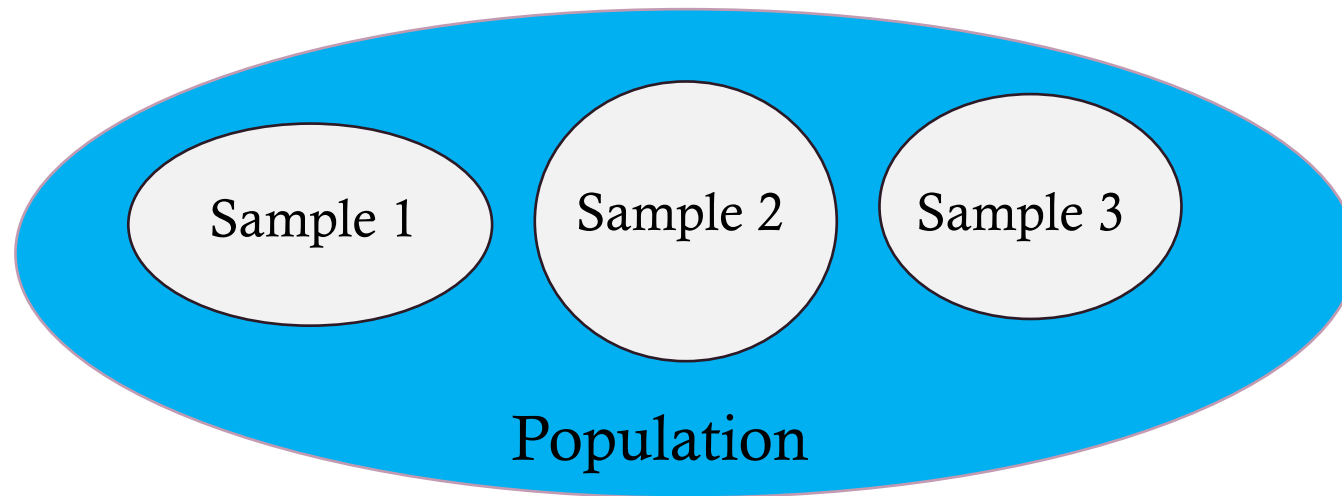


OUTLINE

- Populations, samples and processes
 - Measures of Location
 - Measures of Variability
 - Graphical Displays of Data
-

POPULATION & SAMPLE

- **Statistics** is the science of gathering, describing, and analyzing **data**.
- A **population** is a particular group of interest.
- A **sample** is a subset of the population from which data are collected.



POPULATION & SAMPLE

Example 1: The campus coffee shop of a college wants to know if students have tried the new flavor in their menu. They surveyed 120 randomly chosen students and found that 46 have tried it. Identify the population and the sample for this study.

Population: All students at the college

Sample: 120 randomly chosen students who were surveyed

Example 2: A personnel director is interested in determining how effective a new training course will be in improving the writing skill of her company's employees. The director randomly selects 30 employees and determines their writing skill both before and after taking the training course. Identify the population and the sample for this study.

Population: All employees in the personnel director's company

Sample: 30 employees in the company who were selected to take the new training course

VARIABLE

- A **variable** is a value or characteristic that changes among members of the population.
 - **Data** are the counts, measurements, or observations gathered about a specific variable in a population in order to study it.
 - A **census** is a study in which data are obtained from every member of the population.
-

VARIABLE

Examples: Determine the variable in the following studies.

- 1) A sociologist wishes to estimate the proportion of all adults in a certain region who have never married. In a random sample of 1,320 adults, 145 have never married, hence $145/1320 \approx .11$ or about 11% have never married.
Variable: Marital status (Have never married / Have married)
- 2) After an airplane security scare on Christmas day, 2009, the Gallup organization interviewed 542 American air travelers about increased security measures at airports. The report stated that **78%** of American air travelers are in favor of United States airports using full-body-scan imaging on airline passengers.

Source: Jones, Jeffrey M. "In U.S., Air Travelers Take Body Scans in Stride." 11 Jan. 2010.
<http://www.gallup.com/poll/125018/Air-Travelers-Body-Scans-Stride.aspx> (12 Dec. 2011).

Variable: Opinion about increased security measures at airports (In favor / Against)

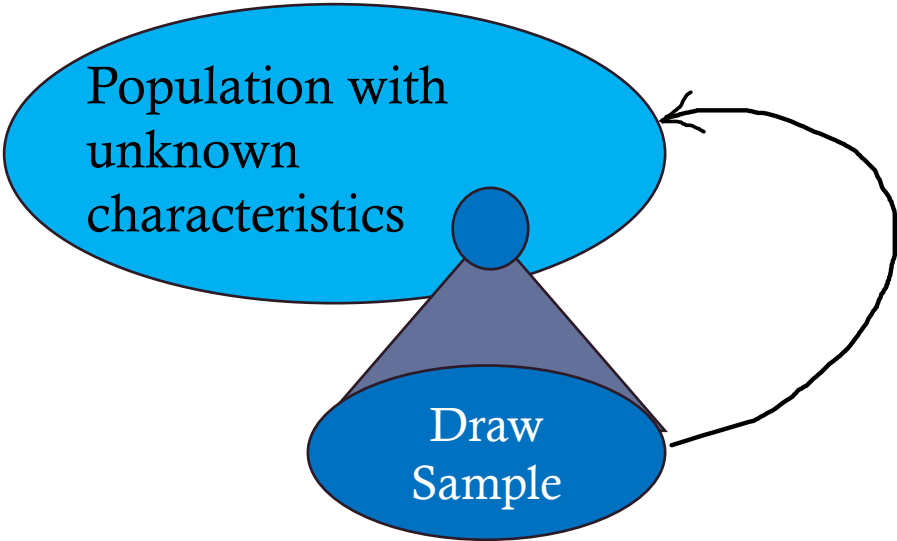
VARIABLE

- A **univariate** data set consists of observations on a single variable.
- A **bivariate** data when observations are made on each of two variables.
- A **multivariate** data arises when observations are made on more than one variable (so bivariate is a special case of multivariate).

Example:

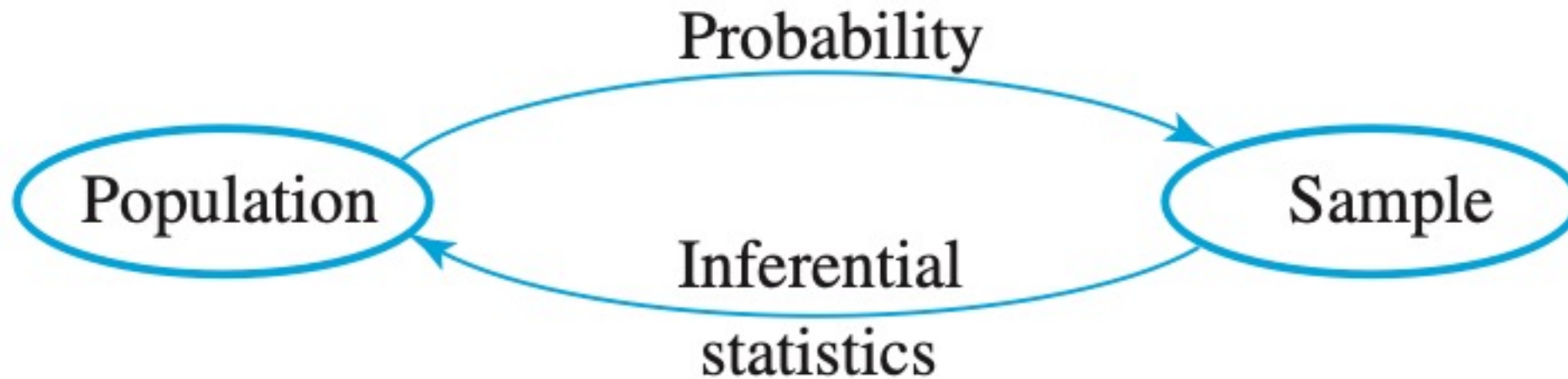
- Physical exam: (height, weight) → bivariate
 - Automobile: (make, model, size, color) → multivariate
-

BRANCHES OF STATISTICS

Descriptive Statistics	Inferential Statistics
<ul style="list-style-type: none">• Gathers, sorts, summarizes, and displays the data• Answers the questions such as:<ol style="list-style-type: none">1. Typical value?2. How much variation?3. Extreme values?4. Shape or distribution of the data?5. Relative position of a value in the data set?6. Any relationships among variables? How strong is the relationship?	<ul style="list-style-type: none">• Makes reasonable guesses about population characteristics using sample data  <pre>graph TD; A([Population with unknown characteristics]) --> B([Draw Sample])</pre>

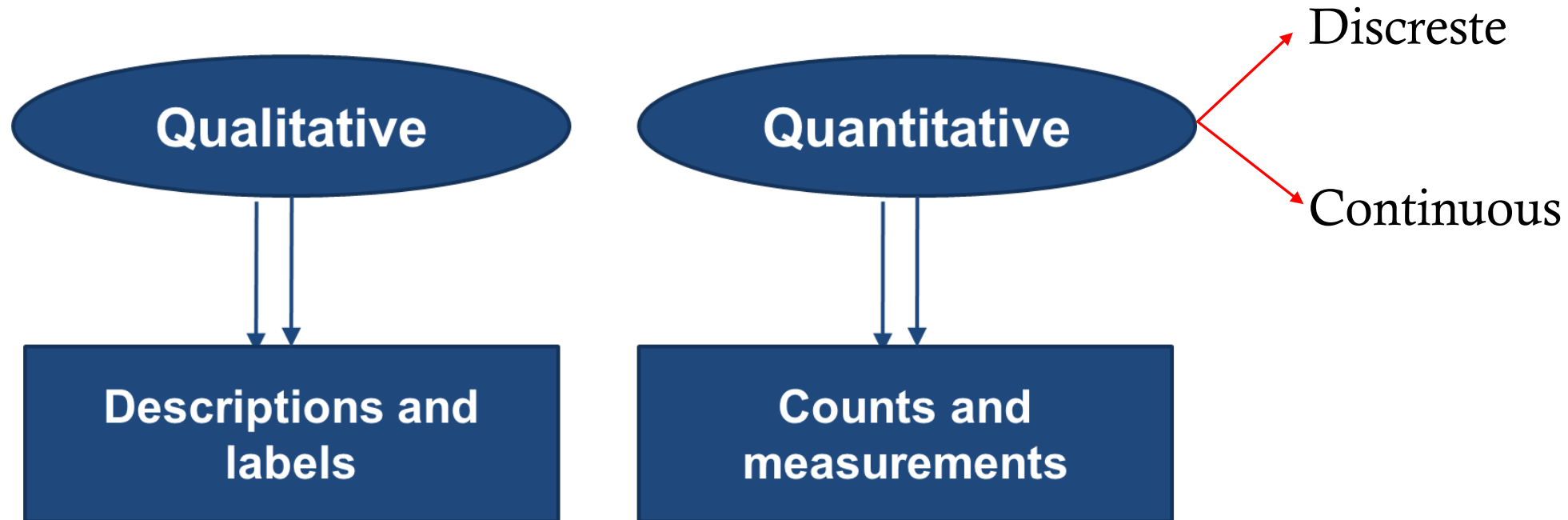
PROBABILITY & STATISTICS

Before we can understand what a particular sample can tell us about the population, we should first understand the uncertainty associated with taking a sample from a given population.



DATA CLASSIFICATION

- Qualitative data, also known as categorical data, consist of labels or descriptions of traits.
- Quantitative data consist of counts and measurements.



DATA CLASSIFICATION

Example: Classify the following data as either qualitative or quantitative.

- a. Shades of red paint in a home improvement store
 - b. Rankings of the most popular paint colors for the season
 - c. Amount of red primary dye necessary to make one gallon of each shade of red paint
 - d. Numbers of paint choices available at several stores
 - e. The genders of the first 40 newborns in a hospital one year.
 - f. The fuel economy in miles per gallon of 20 new cars purchased last month.
-

THE PROCESS OF A STATISTICAL STUDY

Conducting a Statistical Study

1. Determine the design of the study.
 - a) State the question to be studied.
 - b) Determine the population and variables.
 - c) Determine the type of the study, the sampling method.
 2. Collect the data.
 3. Organize the data.
 4. Analyze the data to answer the question.
-

THE PROCESS OF A STATISTICAL STUDY

Example:

1. Does taking 80 mg of aspirin each morning reduce the risk of heart attacks?
Population? Variables? Sampling method?
 2. How to collect data? Observational study / Experiment; Population / Sample;
Sampling Methods?
 3. Organize data: graphical displays; numerical descriptions; compute statistics
 4. Analyze the data to answer the question.
-

ENUMERATIVE VERSUS ANALYTIC STUDIES

Enumerative Study	Analytic Study
<ul style="list-style-type: none">• The goal or purpose of the study is identifiable, i.e., not ambiguous.• The elements of the population are well defined and unchanging under this study.• The population could be an existing finite population about which one wants to draw conclusions. In this discussion, a <i>sampling frame</i>, which is a list of sample points to be collected is either available to an investigator or else can be constructed from definition of the study.• Examples: designed experiment, census	<ul style="list-style-type: none">• Is not an Enumerative study• Draws conclusions about a process that does not even exist at the time of the study. Generally, here the result of the study is new because the objective is to improve things to be used in the future.• The study does not have a well-defined <i>sampling frame</i> and the impact of this study is highly localized and short term.• Example: in the production industry, where the new product is developed as an improvement over the existing one.

OBSERVATIONAL STUDY & EXPERIMENT

- An **observational study** observes data that already exist.
- An **experiment** generates data to be used in the study.

Example:

- 1) You want to determine the average age of college students across the nation.
- 2) Researchers wish to determine if flu shots actually help prevent severe cases of the flu.

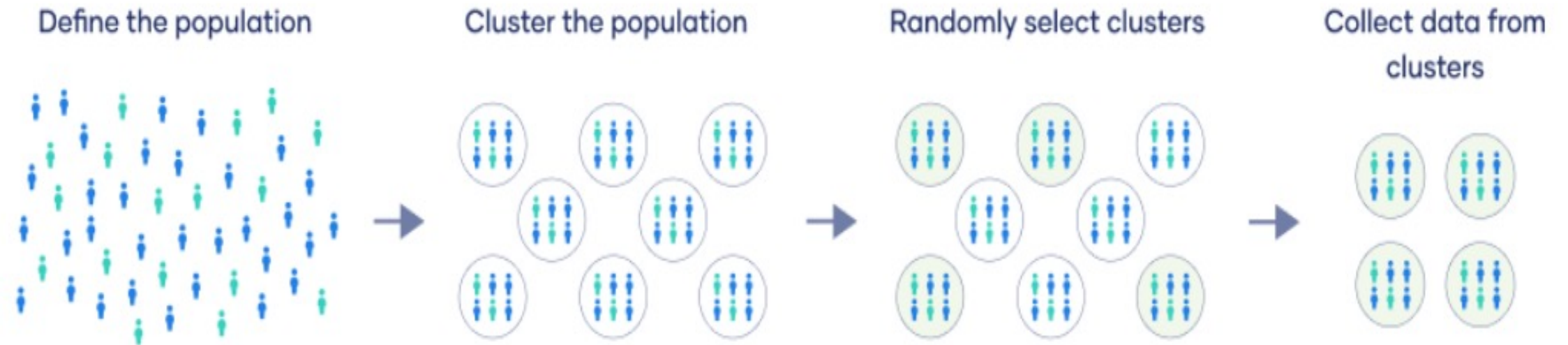
OBSERVATIONAL STUDIES

- A representative sample has the same relevant characteristics as the population and does not favor one group from the population over another.
 - Sample Methods:
 - 1) Random Sampling: Every member of the population has an equal chance of being selected.
 - 2) Simple Random Sampling: Every sample of the population has an equal chance of being selected.
 - 3) Stratified Sampling: A few members from each stratum (or group) are randomly selected.
 - 4) Cluster Sampling: All members from a few randomly chosen clusters (group) are selected.
 - 5) Systematic Sampling: Every n th member of the population is chosen.
 - 6) Convenience Sampling: The sample is chosen because it is convenient for the researcher.
-

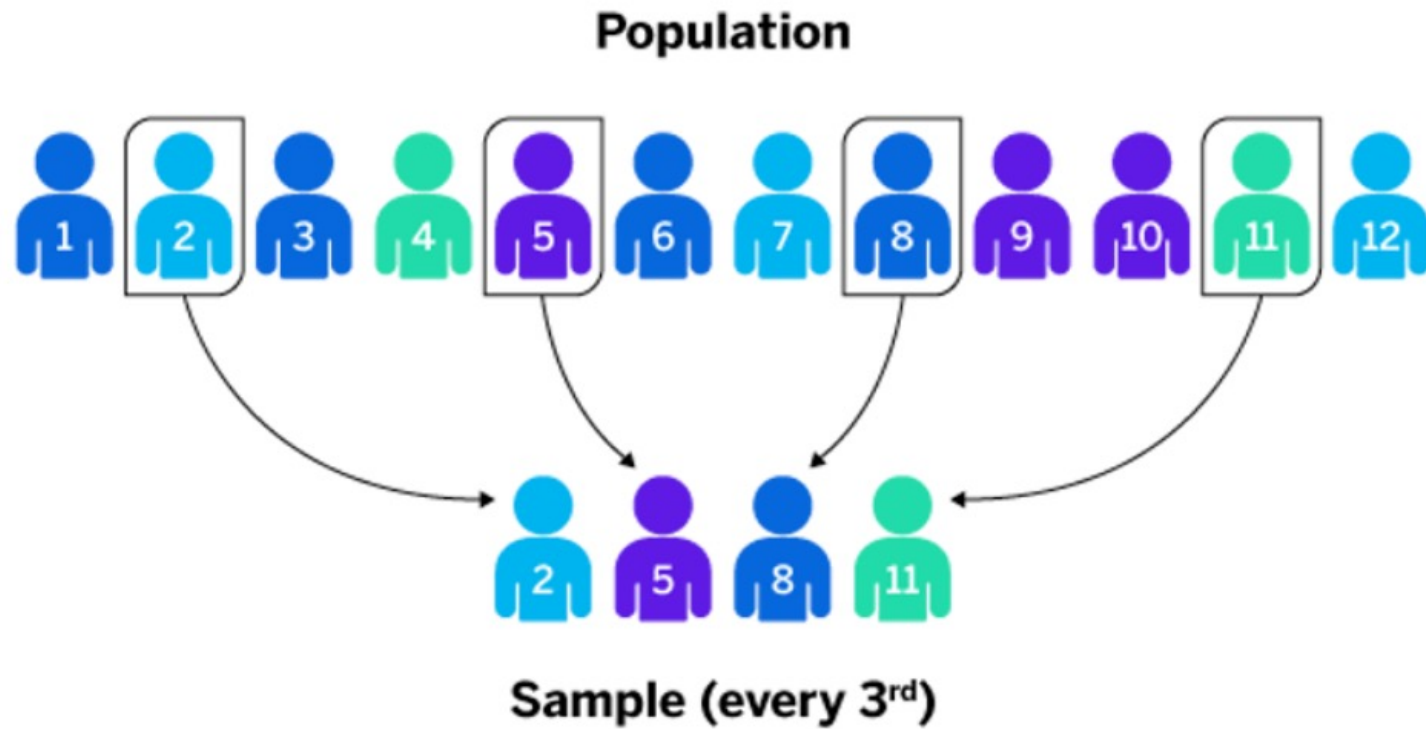
STRATIFIED SAMPLING



CLUSTER SAMPLING



SYSTEMATIC SAMPLING



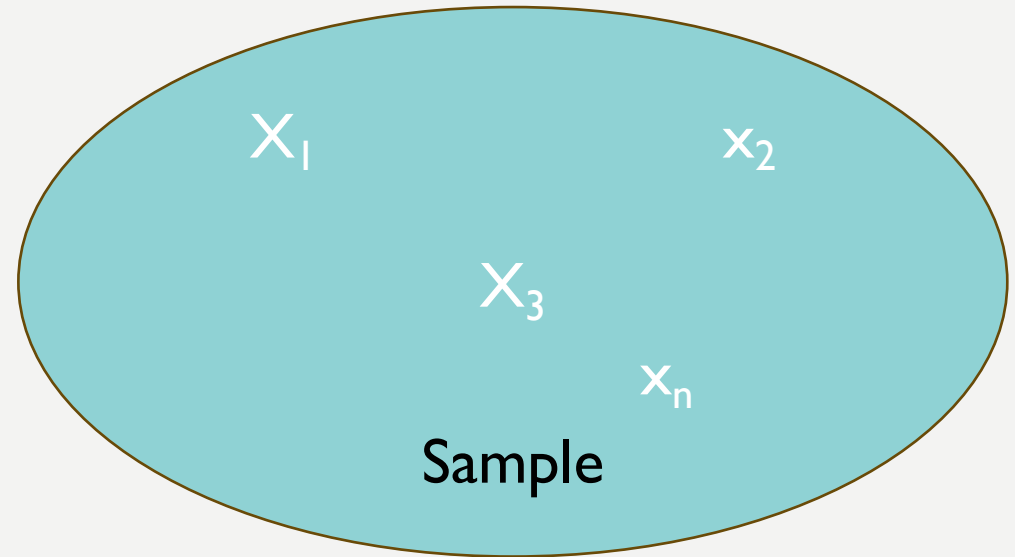
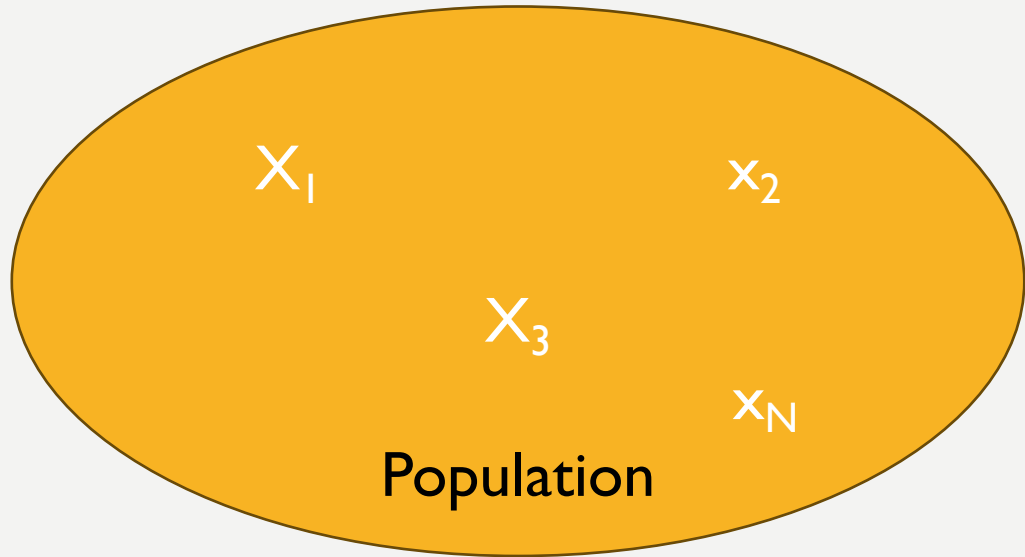
EXPERIMENT

- A **treatment** is some condition that is applied to a group of subjects in an experiment.
 - **Subjects** are people or things being studied in an experiment.
 - **Participants** are people being studied in an experiment.
 - The **response variable** is the variable in an experiment that responds to the treatment.
 - The **explanatory variable** is the variable in an experiment that causes the change in the response variable.
-

MEASURES OF LOCATION

- Mean, median, and mode.
- Quartiles, percentiles, and trimmed means.
- Determine the most appropriate measure of location.

NOTATION



Population

- Population size: N
- The i^{th} data value: x_i
- Population mean: μ

Sample

- Sample size: n
- The i^{th} data value: x_i
- Sample mean: \bar{x}

MEAN

Population Mean

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

$$= \frac{\sum x_i}{N}$$

Sample Mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

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

MEDIAN

Finding the Median of a Data Set

1. List the data in ascending (or descending) order, making an ordered array.
2. If the data set contains an ODD number of values, the median is the middle value in the ordered array.
3. If the data set contains an EVEN number of values, the median is the arithmetic mean of the two middle values in the ordered array. Note that this implies that the median may not be a value in the data set.

MODE

- The **mode** is the value in the data set that occurs most frequently.
- If the data values only occur once or an equal number of times, we say there is **no mode**.
- If one value occurs most often, then the data set is said to be **unimodal**.
- If exactly two values occur equally often, then the data set is said to be **bimodal**.
- If more than two values occur equally often, the data set is said to be **multimodal**.

DETERMINING THE MOST APPROPRIATE MEASURE OF CENTER

1. For qualitative data, the mode should be used.
2. For quantitative data, the mean should be used, unless the data set contains outliers or is skewed.
3. For quantitative data sets that are skewed or contain outliers, the median should be used.

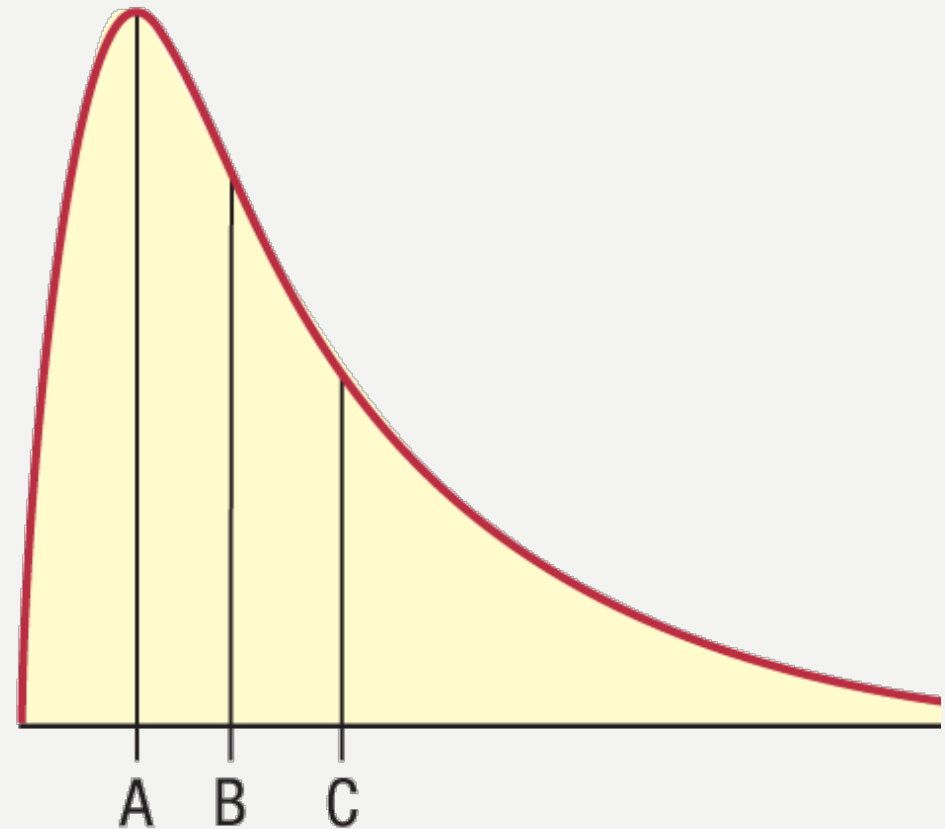
Example:

Choose the best measure of center for the following data sets.

- a. T-shirt sizes (S, M, L, XL) of American women
- b. Salaries for a professional team of baseball players
- c. Prices of homes in a subdivision of similar homes
- d. Rankings of services on a scale of *best*, *average*, and *worst*

GRAPHS AND MEASURES OF CENTER

1. The mode is the data value at which a distribution has its highest peak.
2. The median is the number that divides the area of the distribution in half.
3. The mean of a distribution will be pulled toward any outliers.



PROPERTIES OF MEAN, MEDIAN, AND MODE

Mean	Median	Mode
“Average”	“Middle value”	“Most frequent value”
May not be a data value	May not be a data value	Must be a data value
Single Value	Single Value	Could be one value, multiple values, or not exist
Affected by outliers	Not affected by outliers	Not affected by outliers
Use for quantitative data with <i>no</i> outliers	Use for quantitative data with outliers	Use for qualitative data

TRIMMED MEAN

- A trimmed mean (truncated mean) is a method of **averaging that removes a small percentage of the largest and smallest values** before calculating the mean.
- Using a trimmed mean helps eliminate the influence of outliers or data points on the tails that may unfairly affect the traditional mean.
- Three commonly applied trim percentages: 5%, 10%, and 20%.

To calculate a $X\%$ trimmed mean, you can use the following steps:

- **Step 1:** Order each value in a dataset from smallest to largest.
- **Step 2:** Remove the values in the bottom $X\%$ and top $X\%$ of the dataset.
- **Step 3:** Calculate the mean of the remaining values.

Example:

Given the data set: 4, 8, 12, 15, 9, 6, 14, 18, 12, 9. Calculate the 10% trimmed mean.

Solution:

Ordered Dataset: 4, 6, 8, 9, 9, 12, 12, 14, 15, 18

Trimmed Dataset: 6, 8, 9, 9, 12, 12, 14, 15

$$10\% \text{ trimmed mean} = (6+8+9+9+12+12+14+15) / 8 = 10.625$$

The 10% trimmed mean is **10.625**.

MEASURES OF RELATIVE POSITION

P^{th} Percentile of a Data Value

The P^{th} percentile of a particular value in a data set is given by

$$P = \frac{l}{n} \cdot 100$$

where P is the percentile rounded to the nearest whole number,

l is the number of values in the data set *less than or equal to* the given value, and

n is the number of data values in the sample.

Example:

- a) If the scores of a set of students in a math test 20, 30, 15 and 75, what is the percentile rank of the score 30?

- b) The scores obtained by 10 students are 38, 47, 49, 58, 60, 65, 70, 79, 80, 92. Using the percentile formula, calculate the percentile for score 70?

Location of Data Value for the P^{th} Percentile

To find the *data value* for the P^{th} percentile, the location of the data value in the data set is given by

$$l = n \cdot \frac{P}{100}$$

where l is the location of the P^{th} percentile in the *ordered array* of data values.

n is the number of data values in the sample, and P stands for the P^{th} percentile.

- If the formula results in a decimal value for l , the location is the next *larger* whole number. (Note: Other methods may be used here.)

Example:

Consider the data set {50, 45, 60, 25, 30}. Find the 5th, 30th, 40th, 50th and 100th percentiles of the list given.

Solution:

Ordered list – 25, 30, 45, 50, 60
N = 5

Percentile (P)	Ordinal rank	Percentile value
5th	$(5/100) \times 5 = [0.25] = 1$	1st number in the ordered list = 25
30th	$(30/100) \times 5 = [1.5] = 2$	2nd number in the ordered list = 30
40th	$(40/100) \times 5 = 2$	2nd number in the ordered list = 30
50th	$(50/100) \times 5 = [2.5] = 3$	3rd number in the ordered list = 45
100th	$(100/100) \times 5 = 5$	5th number in the ordered list = 60

Quartiles

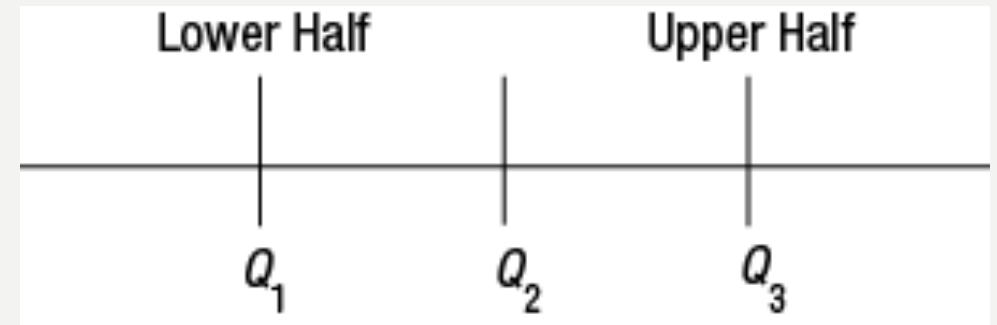
Q_1 = First Quartile: 25% of the data are less than or equal to this value.

Q_2 = Second Quartile: 50% of the data are less than or equal to this value.

Q_3 = Third Quartile: 75% of the data are less than or equal to this value.

Five-number summary

- The minimum value
- The first quartile, Q_1
- The median, or second quartile, Q_2
- The third quartiles, Q_3
- The maximum value



Interquartile Range: $IQR = Q_3 - Q_1$

BOX PLOT

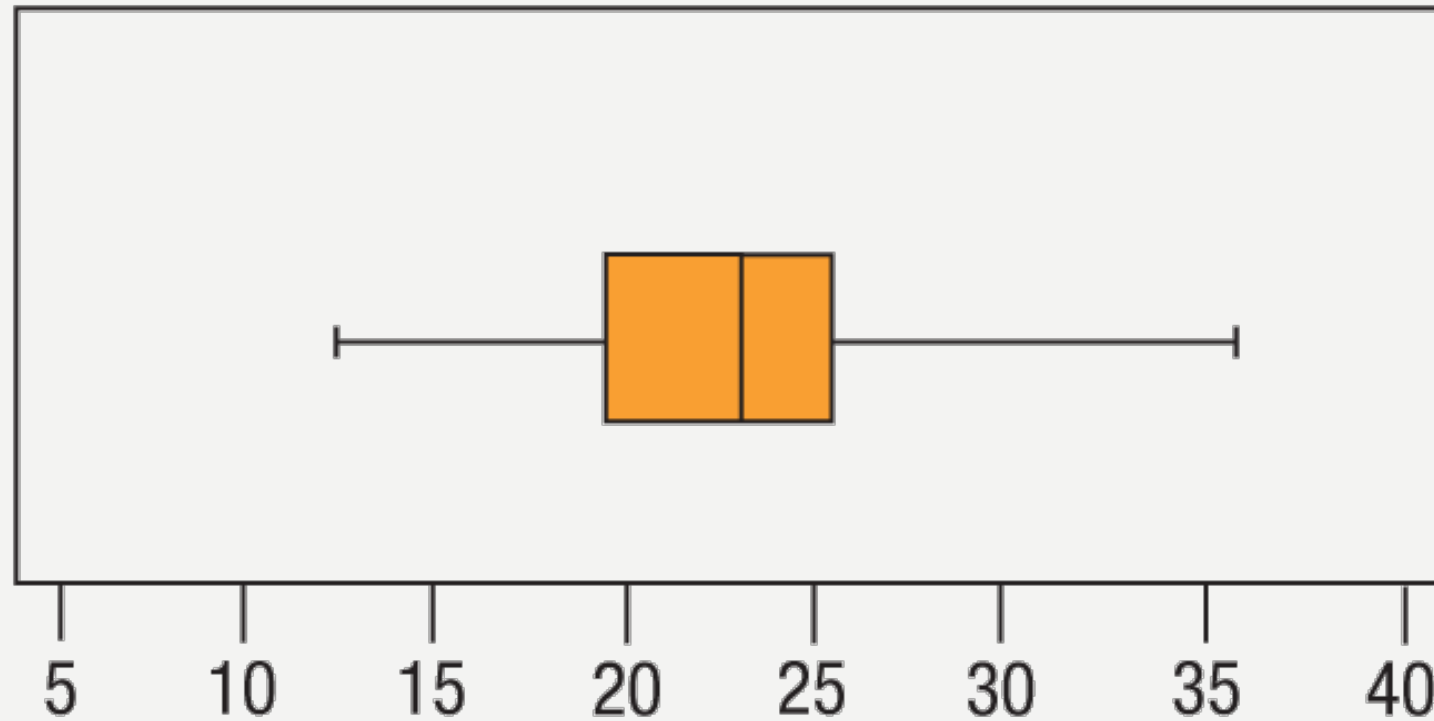
Creating a Box Plot

1. Begin with a horizontal (or vertical) number line that contains the five-number summary.
2. Draw a small line segment above (or next to) the number line to represent each of the numbers in the five-number summary.
3. Connect the line segment that represents the first quartile to the line segment representing the third quartile, forming a box with the median's line segment in the middle.
4. Connect the "box" to the line segments representing the minimum and maximum values to form the "whiskers."

Example:

Draw a box plot to represent the five-number summary:

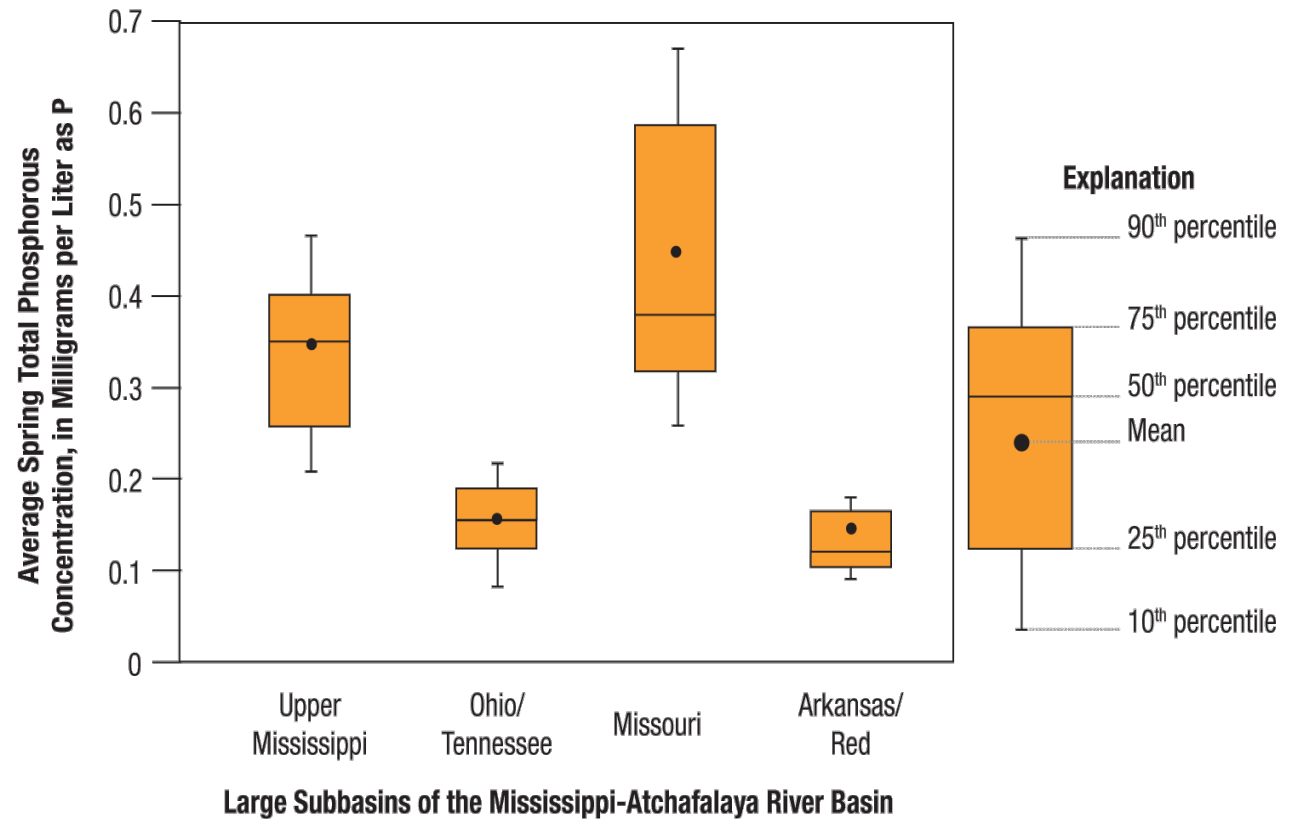
12.1, 19.8, 23.6, 25.3, 35.9



Example:

The box plots below are from the US Geological Survey website. Use them to answer the following questions.

- What do the top and bottom bars represent in these box plots according to the key?
- Which subbasin had the highest median average spring total phosphorus concentration?
- Which subbasin had the lowest average spring total phosphorus concentration? (**Note:** Each data value is an average of April's and May's totals, and the lowest average shown for each subbasin is the 10th percentile.)
- Which subbasin had the largest interquartile range?





MEASURES OF VARIABILITY

- Range, variance, and standard deviation.
-

Range

$$\text{Range} = \text{Maximum Data Value} - \text{Minimum Data Value}$$

Example:

The following data were collected from samples of call lengths (in minutes) observed for two different mobile phone users. Calculate the range of each data set.

- a. 2, 25, 31, 44, 29, 14, 22, 11, 40
- b. 2, 2, 44, 2, 2, 2, 2, 2
- c. What could be misleading about using the range as a measurement?

Variance

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$ <ul style="list-style-type: none">• x_i is the i^{th} value in the population• μ is the population mean• N is the number of values in the population	$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$ <ul style="list-style-type: none">• x_i is the i^{th} data value• \bar{x} is the sample mean• n is the number of data values in the sample

Standard Deviation

- The **standard deviation** is a measure of how much we might expect a typical member of the data set to differ from the mean. It is the **square root of the variance**.

Population Standard Deviation	Sample Standard Deviation
$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$

Note:
$$\sum (x_i - \bar{x})^2 = \sum (x_i)^2 - \frac{(\sum x_i)^2}{n}$$

Properties of the Standard Deviation

- s measures spread about the mean. Use s to describe the spread of a distribution only when you use the mean to describe the center.
- $s = 0$ only when there is no spread. This happens only when all observations have the same value. So standard deviation zero means no spread at all. Otherwise, $s > 0$. As the observations become more spread out about their mean, s gets larger.

GRAPHICAL DISPLAYS OF DATA

- Histogram, stem-and-leaf plot, dot plot, scatter plot.

Data Frequency Table

- A **data frequency table**, a table in which each distinct value x is listed in the first row and its frequency f , which is the number of times the value x appears in the data set, is listed below it in the second row.

The data set of the previous example is represented by the data frequency table:

x	17	18	19	20	21	22	24
f	2	8	5	3	1	1	1

Example: Creating a Stem-and-Leaf Plot

- Create a stem-and-leaf plot of the following ACT scores from a group of college freshmen.

ACT Scores				
18	23	24	31	19
27	26	22	32	18
35	27	29	24	20
18	17	21	25	26

ACT Scores

Stem	Leaves
1	8 9 8 8 7
2	3 4 7 6 2 7 9 4 0 1 5 6
3	1 2 5

Key: 1|8 = 18

ACT Scores

Stem	Leaves
1	7 8 8 8 9
2	0 1 2 3 4 4 5 6 6 7 7 9
3	1 2 5

Key: 1|8 = 18

Stem-and-Leaf Plots

A **stem-and-leaf plot** is a graph of quantitative data that visually displays the data.

Characteristics:

- A stem-and-leaf plot retains the original data.
- The leaves are usually the last digit in each data value and the stems are the remaining digits.
- A legend, sometimes called a **key**, should be included so that the reader can interpret the information.

Stem-and-Leaf Plots

Constructing a Stem-and-Leaf Plot

1. Create two columns, one on the left for stems and one on the right for leaves.

List each stem that occurs in the data set in numerical order. Each stem is normally listed only once; however, the stems are sometimes listed two or more times if splitting the leaves would make the data set's features clearer.

3. List each leaf next to its stem. Each leaf will be listed as many times as it occurs in the original data set. There should be as many leaves as there are data values. Be sure to line up the leaves in straight columns so that the table is visually accurate.
4. Create a key to guide interpretation of the stem-and-leaf plot.
5. If desired, put the leaves in numerical order to create an **ordered stem-and-leaf plot**.

Example: Creating and Interpreting a Stem-and-Leaf Plot

Create a stem-and-leaf plot for the following starting salaries for entry-level accountants at public accounting firms. Use the stem-and-leaf plot that you create to answer the following questions.

Starting Salaries for Entry-Level Accountants				
\$51,500	\$48,300	\$40,900	\$40,700	\$48,200
\$45,500	\$42,500	\$44,200	\$46,400	\$48,600
\$45,800	\$46,300	\$50,000	\$50,700	\$44,300
\$43,000	\$42,700	\$49,000	\$46,700	\$43,200
\$42,900	\$46,500	\$47,700	\$48,000	\$46,300
\$44,500	\$47,900	\$45,300	\$46,100	\$45,000

- What were the smallest and largest salaries recorded?
- Which salary appears the most often?
- How many salaries were in the range \$41,000–\$41,900?
- In which salary range did the most salaries lie: \$40,000–\$44,900, \$45,000–\$49,900, or \$50,000 and above?

Example: Creating and Interpreting a Stem-and-Leaf Plot (cont.)

Starting Salaries for Entry-Level Accountants

Stem	Leaves
40	7 9
41	
42	5 7 9
43	0 2
44	2 3 5
45	0 3 5 8
46	1 3 3 4 5 7
47	7 9
48	0 2 3 6
49	0
50	0 7
51	5

Key: 40 | 7 = \$40,700

Histograms

A **frequency histogram**, shortened to histogram, is a bar graph of a frequency distribution of quantitative data.

A **relative frequency histogram** is a histogram in which the heights of the bars represent the relative frequencies of each class rather than simply the frequencies.

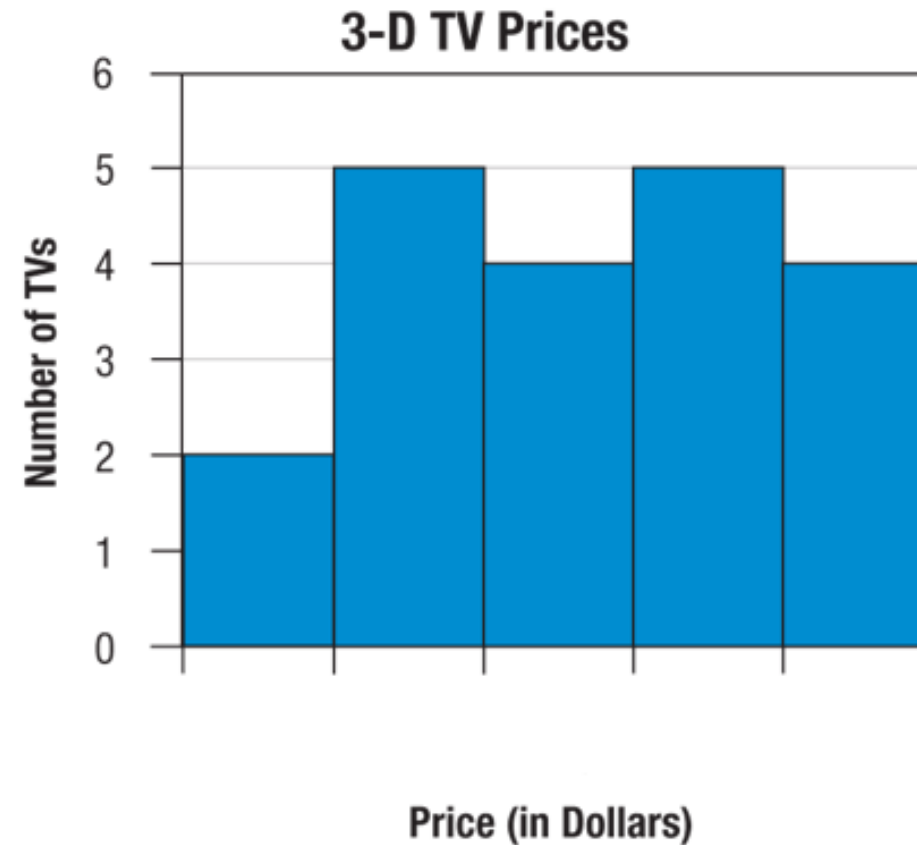
Characteristics of Histograms

- A bar graph of a frequency distribution.
- The horizontal axis is a real number line.
- The width of the bars represent the class width from the frequency table and should be uniform.
- The bars in a histogram should touch.
- The height of each bar represents the frequency of the class it represents.

Example: Constructing a Histogram

Construct a histogram of the 3-D TV prices from the previous section. The frequency distribution of the data is restated here.

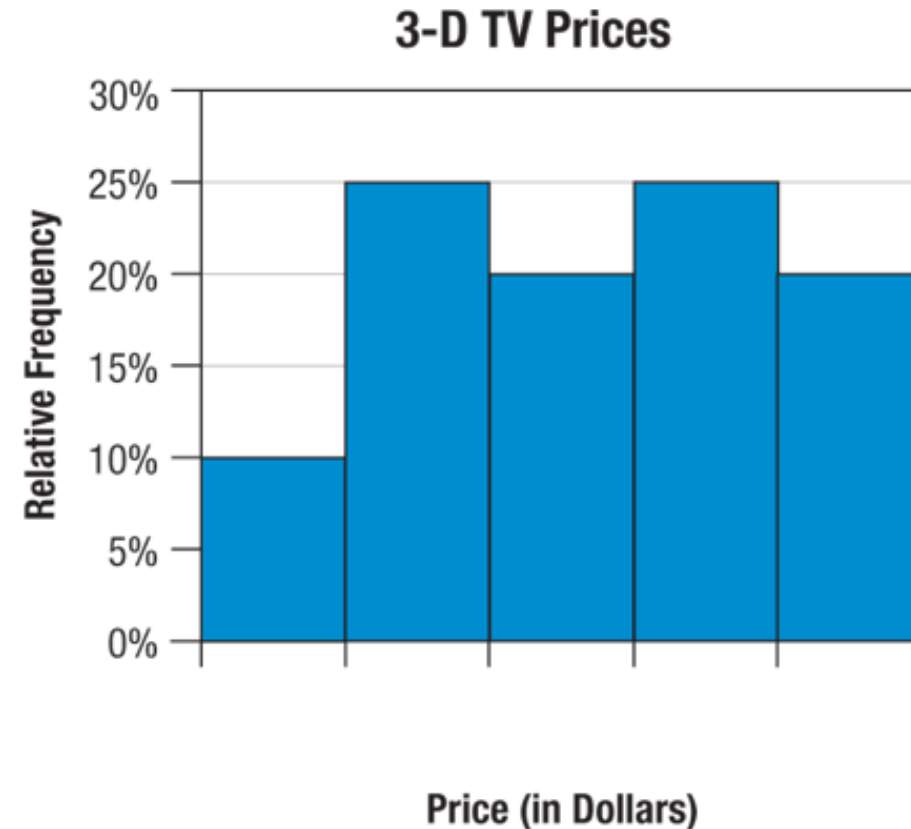
3-D TV Prices	
Class	Frequency
\$1500 - \$1599	2
\$1600 - \$1699	5
\$1700 - \$1799	4
\$1800 - \$1899	5
\$1900 - \$1999	4



Example: Constructing a Relative Frequency Histogram

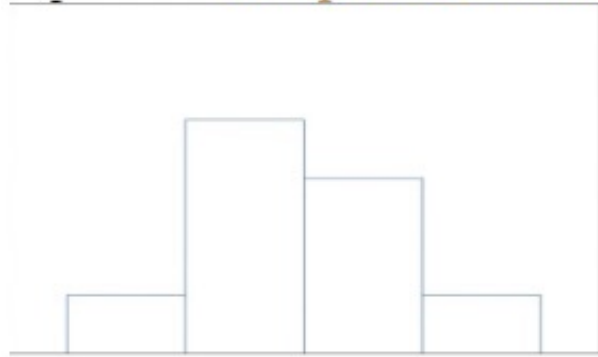
Construct a relative frequency histogram of the 3-D TV prices from the previous example. The frequency distribution of the data is reprinted here.

3-D TV Prices		
Class	Frequency	Relative Frequency
\$1500 - \$1599	2	
\$1600 - \$1699	5	
\$1700 - \$1799	4	
\$1800 - \$1899	5	
\$1900 - \$1999	4	



Shapes of Histograms

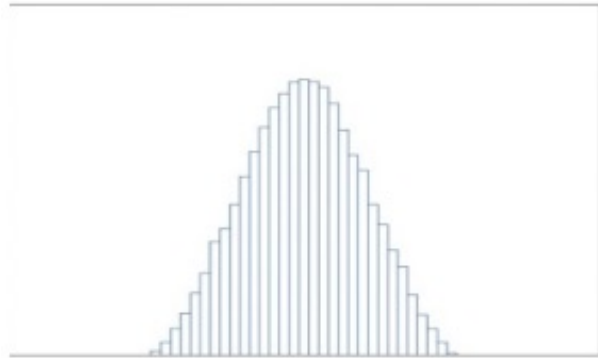
Sample Size and Relative Frequency Histograms



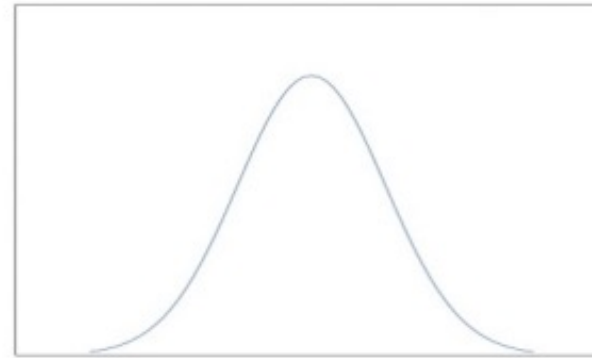
(a) Small Sample



(b) Medium Sample



(c) Large Sample



(d) Very Large Sample

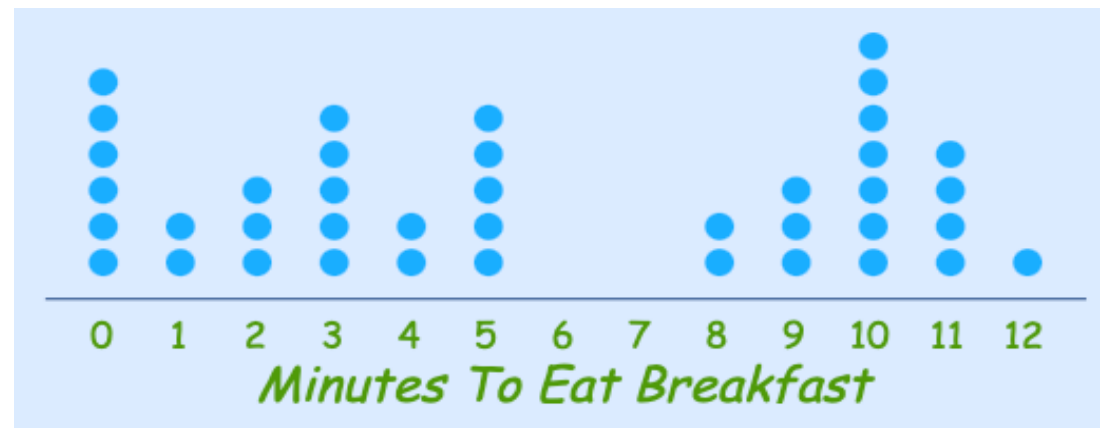
Dot Plot

- A dotplot is an attractive summary of numerical data when the data set is reasonably small or there are relatively few distinct data values.

Example:

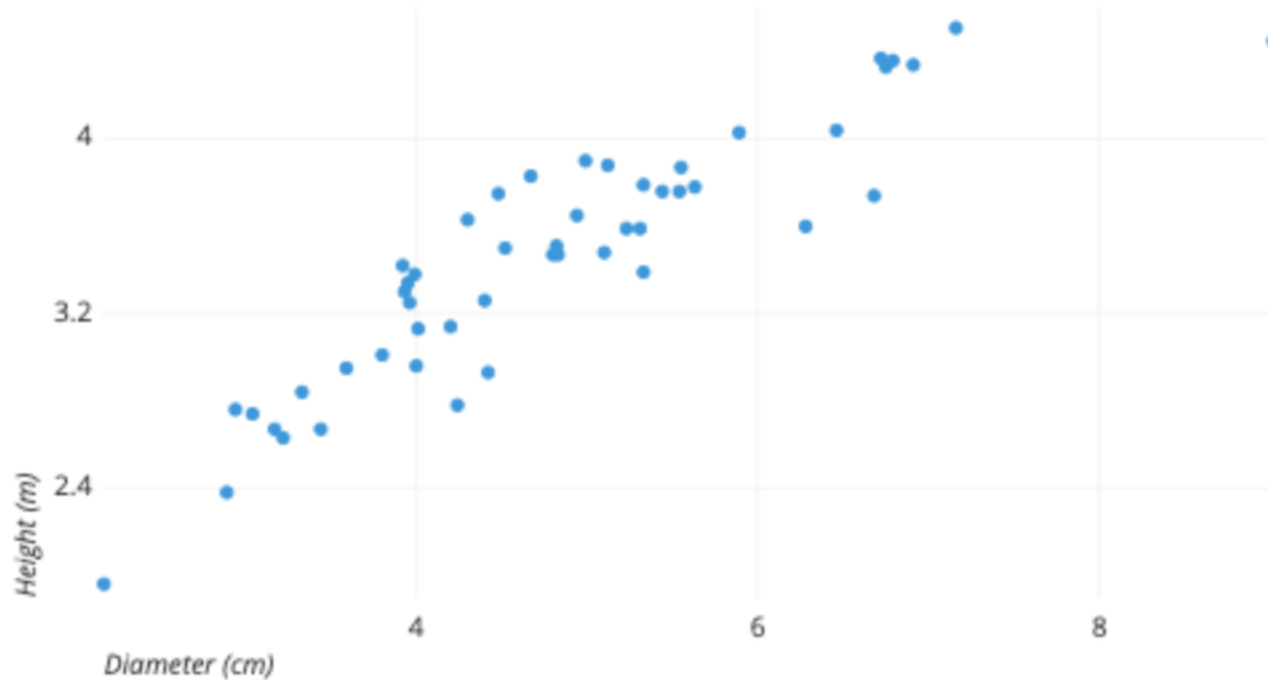
A survey of "How long does it take you to eat breakfast?" has these results:

Minutes	1	2	3	4	5	6	7	8	9	10	11	12
People	6	2	3	5	2	5	0	0	2	3	7	4



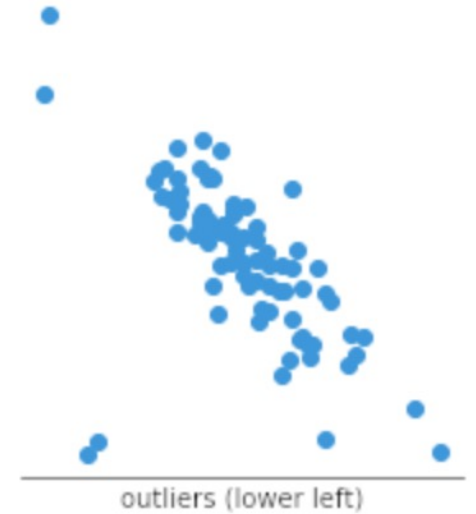
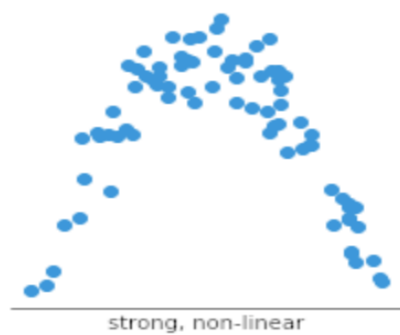
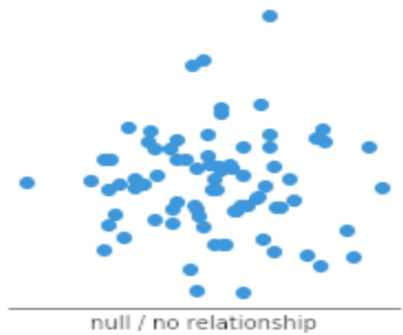
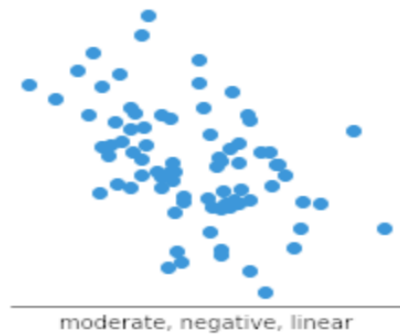
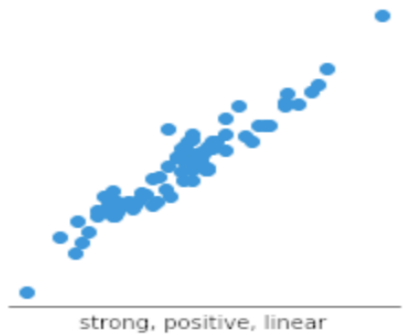
Scatter Plot

- A scatter plot uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables.



Scatter Plot

- A scatter plot is useful for identifying patterns in data. Note that correlation does not imply causation.



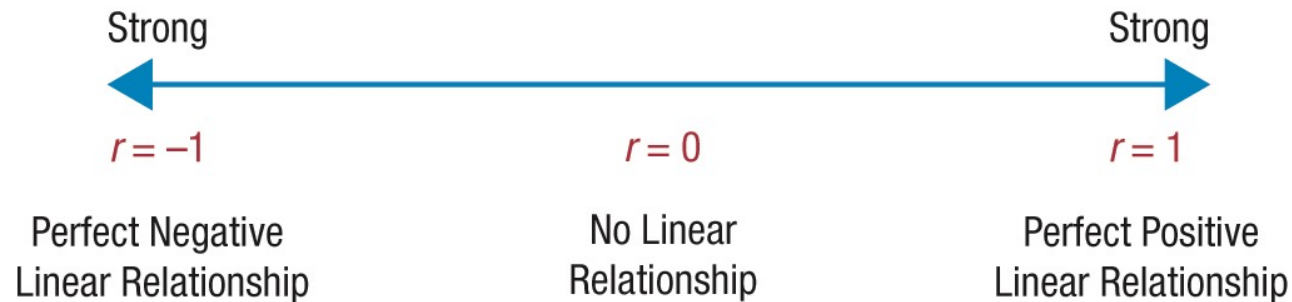
Scatter Plots and Correlation

- The **Pearson correlation coefficient**, ρ , is the parameter that measures the strength of a linear relationship between two quantitative variables in a population. The correlation coefficient for a sample is denoted by r .

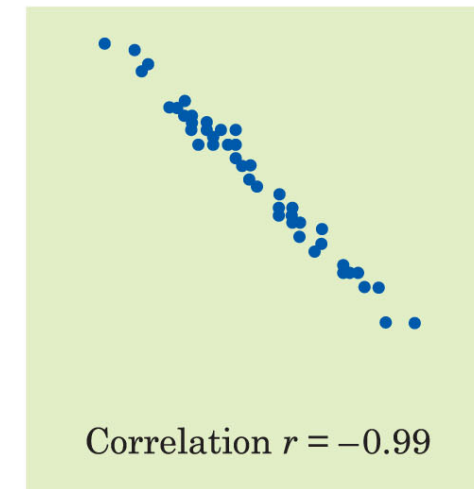
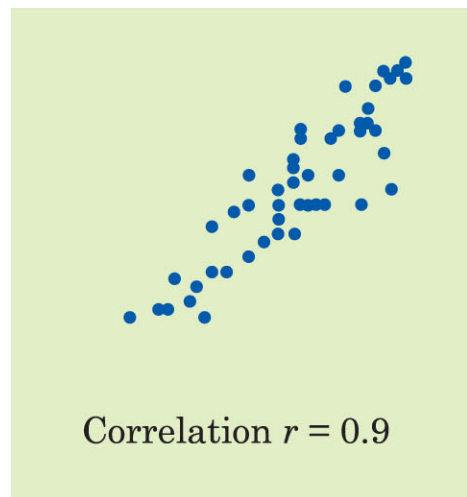
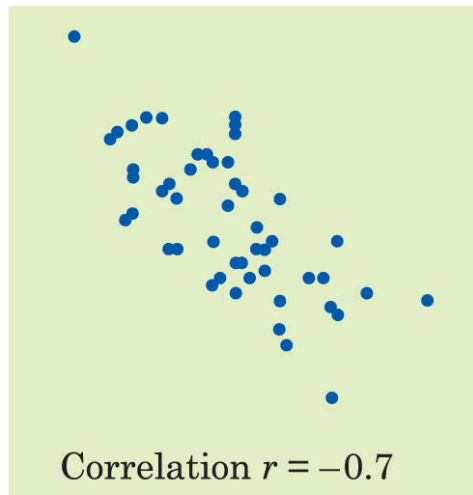
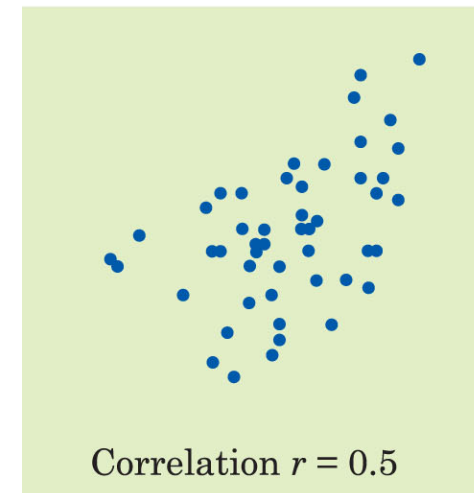
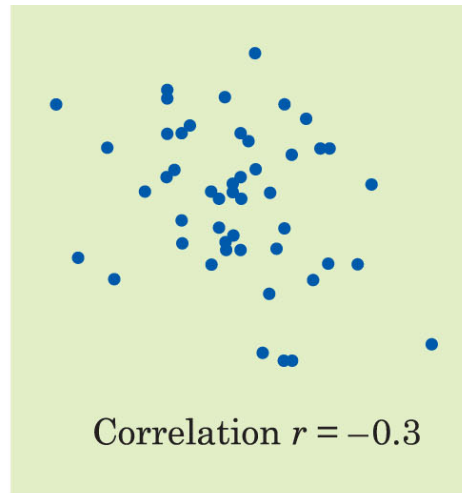
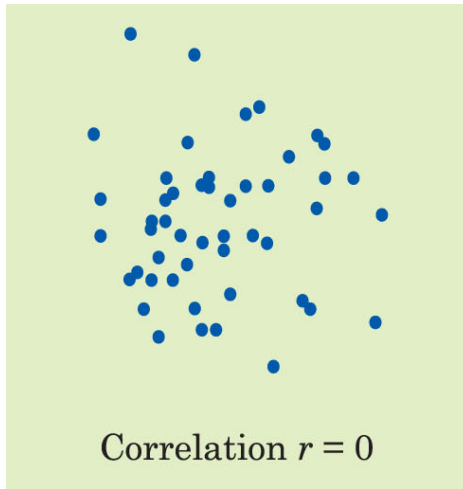
$$-1 \leq r \leq 1$$

- If two variables have a strong positive or negative relationship, we say that the two variables are correlated. The strength of the correlation is expressed by $|r|$. The larger $|r|$ is, the stronger the correlation.

Pearson Correlation Coefficient, r



Scatterplots and Correlation



Scatter Plots and Correlation

Pearson Correlation Coefficient

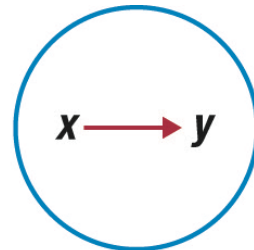
The **Pearson (linear) correlation coefficient** for paired data from a sample is given by

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}} \cdot \sqrt{SS_{yy}}} = \frac{\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)}{\sqrt{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} \sqrt{\sum y_i^2 - \frac{1}{n}(\sum y_i)^2}}$$

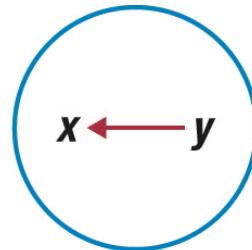
where n is the number of data pairs in the sample, x_i is the i^{th} value of the explanatory variable, and y_i is the i^{th} value of the response variable.

Statistics and Causation

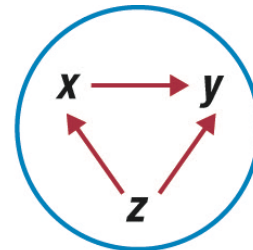
1. A strong relationship between two variables does not always mean that changes in one variable cause changes in the other.
2. The relationship between two variables is often influenced by other variables lurking in the background.
3. The best evidence for causation comes from randomized comparative experiments.



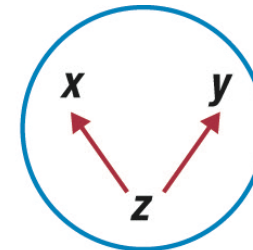
Variable x
causes y



Variable y
causes x



Variable x influences y ,
but both are also
influenced by z



Variable x does not
cause y , but both are
influenced by z

Practice Problems

1) Suppose a gymnast in the London Olympics received the following scores.

7.5 8 9.5 6.5 7 7.5 8 7.5 8 7

Calculate the trimmed mean by trimming 10% of the data.

2) There are 25 test scores such as: 72, 54, 56, 61, 62, 66, 68, 43, 69, 69, 70, 71, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99. Find the 60th percentile?

3) In a college, a list of grades of 15 students has been declared. Their grades are given as: 85, 34, 42, 51, 84, 86, 78, 85, 87, 69, 74, 65. Find the 80th percentile?

Practice Problems

4) In a college, a list of scores of 10 students is announced. The scores are 56, 45, 69, 78, 72, 94, 82, 80, 63, 59. Using the percentile formula, find the 70th percentile.

5) Find the 85th percentile score in the following test results.

{95, 88, 70, 75, 83, 70, 66, 91, 68, 76, 82}

6) Compute the five number summary for the following data:

4, 8, 11, 11, 12, 14, 16, 20, 21, 25

7) Compute the five number summary for the following data:

1, 1, 1, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 6, 7, 8, 9, 10, 15

8) Create a box plot for the data sets in problem 8) and 9).