

# Lecture One

## Section 1.1 – Introduction to the Practice Statistical

### Define statistics and statistical thinking

What is statistics? Many people say that statistics is numbers. After all, we are bombarded by numbers that supposedly represent how we feel and who we are. For example, we hear on the radio that 50% of first marriages, 67% of second marriages, and 74% of third marriages end in divorce.

Certainly, statistics has a lot to do with numbers, but this definition is only partially correct. Statistics is also about where the numbers come from and how closely the numbers reflect reality.

### Definitions

**Data** are collections of observations (such as measurements, genders, survey responses) and are a “fact or proposition used to draw a conclusion or make a decision.” Data describe characteristics of an individual. A key aspect of data is that they vary. Is everyone in your class the same height? No! Does everyone have the same hair color? No! So, among individuals there is variability.

**Statistics** is the science of planning studies and experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on the data. In addition, statistics is about providing a measure of confidence in any conclusions.

A **population** is the complete collection of all individuals (scores, people, measurements, and so on) to be studied. The collection is complete in the sense that it includes all of the individuals to be studied.

A **census** is the collection of data from every member of the population.

A **sample** is a subcollection of members selected from a population.

One goal of statistics is to describe and understand sources of variability.




### Explain the Process of Statistics

### Definitions

The entire group of individuals to be studied is called the **population**. An **individual** is a person or object that is a member of the population being studied. A **sample** is a subset of the population that is being studied.

### Key Concept

The subject of statistics is largely about using sample data to make inferences (or generalizations) about an entire population. It is essential to know and understand the definitions that follow.

<i>Population</i>	<i>Sample</i>	<i>Individual</i>
		

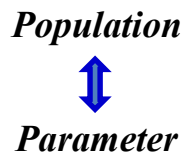
## ***Definitions***

***Descriptive statistics*** consist of organizing and summarizing data. Descriptive statistics describe data through numerical summaries, tables, and graphs. A ***statistic*** is a numerical summary based on a sample.

***Inferential statistics*** uses methods that take results from a sample, extends them to the population, and measures the reliability of the result.



A ***parameter*** is a numerical measurement describing some characteristic of a population.



## ***Example***

There are exactly 100 Senators in the 109<sup>th</sup> Congress of the US, and 55% of them are Republicans. The figure of 55% is a ***parameter*** because it is based on the entire population of all 100 Senators.

## ***Example***

In 1936, *Literary Digest* polled 2.3 million adults in the US, and 57% said that they would vote for Alf Landon for the presidency. That figure of 57% is a ***statistic*** because it is based on a sample, not the entire population of all adults in the US.

## ***Example***

Suppose the percentage of all students on your campus who own a car is 48.1%. This value represents a ***parameter*** because it is a numerical summary of a population.

Suppose a sample of 90 students is obtained, and from this sample we find that 42.5% have a job. This value represents a ***statistic*** because it is a numerical summary based on a sample.

## The Process of Statistics

1. *Identify the research objective:* To determine whether males accused of battering their intimate female partners that were assigned into a 40-hour batter treatment program are less likely to batter again compared to those assigned to 40-hours of community service.
2. *Collect the information needed to answer the question:* The researchers randomly divided the subjects into two groups. Group 1 participants received the 40-hour batterer program, while group 2 participants received 40 hours of community service. Six months after the program ended, the percentage of males that battered their intimate female partner was determined.
3. *Describe the data - Organize and summarize the information:* The demographic characteristics of the subjects in the experimental and control group were similar. After the six month treatment, 21% of the males in the control group had any further battering incidents, while 10% of the males in the treatment group had any further battering incidents.
4. *Draw conclusions from the data:* We extend the results of the 376 males in the study to all males who batter their intimate female partner. That is, males who batter their female partner and participate in a batter treatment program are less likely to batter again.

### Example

Many studies evaluate batterer treatment programs, but there are few experiments designed to compare batterer treatment programs to non-therapeutic treatments, such as community service. Researchers designed an experiment in which 376 male criminal court defendants who were accused of assaulting their intimate female partners were randomly assigned into either a treatment group or a control group. The subjects in the treatment group entered a 40-hour batterer treatment program while the subjects in the control group received 40 hours of community service. After 6 months, it was reported that 21% of the males in the control group had further battering incidents, while 10% of the males in the treatment group had further battering incidents. The researchers concluded that the treatment was effective in reducing repeat battering offenses.

## Distinguish between Qualitative and Quantitative Variables

**Variables** are the characteristics of the individuals within the population.

**Key Point:** Variables vary. Consider the variable height. If all individuals had the same height, then obtaining the height of one individual would be sufficient in knowing the heights of all individuals. Of course, this is not the case. As researchers, we wish to identify the factors that influence variability.

### Definitions

**Qualitative or Categorical variables** allow for classification of individuals based on some attribute or characteristic.

**Example:** The genders (male/female) of professional athletes

**Example:** Shirt numbers on professional athletes uniforms - substitutes for names.

**Quantitative variables** provide numerical measures of individuals. Arithmetic operations such as addition and subtraction can be performed on the values of the quantitative variable and provide meaningful results.

**Example:** The weights of supermodels

**Example:** The ages (in years) of survey respondents

### **Example**

Determine whether the following variables are qualitative or quantitative

- a) Gender
- b) Temperature
- c) Number of days during the past week that a college student studied
- d) Zip code
- e) Nationality
- f) Number of children
- g) Household income in the previous year
- h) Level of education
- i) Daily intake of whole grains (measured in grams per day)

### **Solution**

- a) Gender is a **qualitative** variable because it allows a researcher to categorize the individual as male or female. (Notice that arithmetic operations cannot be performed on these attributes.)
- b) Temperature is a **quantitative** variable because it is numeric, and operations such as addition and subtraction provide meaningful results.
- c) Number of days during the past week that a college student studied is a **quantitative** variable because it is numeric, and operations such as addition and subtraction provide meaningful results.
- d) Zip code is a **qualitative** variable because it categorizes a location. Notice that, even though they are numeric, adding and subtracting zip codes does not provide meaningful results.
- e) Nationality **Qualitative**
- f) Number of children **Quantitative**
- g) Household income in the previous year **Quantitative**
- h) Level of education **Qualitative**
- i) Daily intake of whole grains (measured in grams per day) **Quantitative**

## Distinguish between Discrete and Continuous Variables

Quantitative data can further be described by distinguishing between *discrete* and *continuous* types.

### Definitions

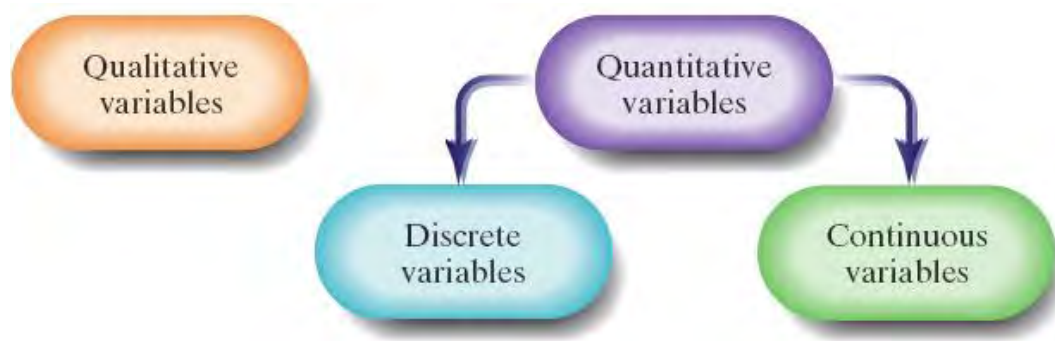
A **discrete variable** is a quantitative variable that has either a finite number of possible values or a countable number of possible values. The term “countable” means the values result from counting such as 0, 1, 2, 3, and so on.

**Example** The number of eggs that hens lay are *discrete* data because they represent counts.

A **continuous variable** is a quantitative variable that has an infinite number of possible values it can take on and can be measured to any desired level of accuracy. The result from infinitely many possible values that correspond to some continuous scale that covers a range of values without gaps, interruptions, or jumps

### Example

The amounts of milk from cows are continuous data because they are measurements that can assume any value over a continuous span. During a year, a cow might yield an amount of milk that can be any value between 0 and 7000 liters. It would be possible to get 2.343115 gallons per day



### Example

Classify each of the following quantitative variables considered in the study as discrete or continuous.

- a) The number of heads obtained after flipping a coin five times.
- b) The number of cars that arrive at a restaurant drive thru between 1:00 PM and 2:00 PM
- c) The distance a car can travel in city driving conditions with a full tank of gas.
- d) Number of children
- e) Household income in the previous year
- f) Daily intake of whole grains (measured in grams per day)

### Solution

- a) The number of heads obtained after flipping a coin five times is a **discrete** variable because we can count the number of heads obtained. The possible values of this discrete variable are 0, 1, 2, 3, 4, 5.

- b) The number of cars that arrive at a restaurant drive thru between 1:00 PM and 2:00 PM is a **discrete** variable because we find its value by counting the cars. The possible values of this discrete variable are 0, 1, 2, 3, 4, and so on.
- c) The distance a car can travel in city driving conditions with a full tank of gas is a **continuous** variable because we can measure the distance.
- d) Number of children is a **discrete** variable
- e) Household income in the previous year is a **continuous** variable
- f) Daily intake of whole grains (measured in grams per day) is a **continuous** variable

## Definitions

The list of observations a variable assumes is called **data**.

While gender is a variable, the observations, male or female, are data.

**Qualitative data** are observations corresponding to a qualitative variable.

**Quantitative data** are observations corresponding to a quantitative variable.

- **Discrete data** are observations corresponding to a discrete variable.
- **Continuous data** are observations corresponding to a continuous variable.

## Example

The table below represents group of selected countries and information regarding these countries. Identify the individuals, variables, and data

<i>Country</i>	<i>Government Type</i>	<i>Life Expectancy (years)</i>	<i>Population (in Millions)</i>
Australia	Federal parliamentary democracy	81.63	21.3
Canada	Constitutional monarchy	81.23	33.5
France	Republic	80.98	64.4
Morocco	Constitutional monarchy	75.47	31.3
Poland	Republic	75.63	38.5
Sri Lanka	Republic	75.14	21.3
United States	Federal republic	78.11	307.2

## Solution

The **individuals** in the study are the countries: Australia, Canada, and so on.

The variables measured for each country are government type, life expectancy, and population. The variable government type is qualitative because it categorizes the individual. The variables life expectancy and population are quantitative.

The quantitative variable life expectancy is continuous because it is measured. The quantitative variable population is discrete because we count people. The **observations** are the data. For example, the data corresponding to the variable life expectancy are 81.63, 81.23, 80.98, 75.47, 75.63, 75.14, and 78.11. The following data correspond to the individual Poland: a republic government with residents whose life expectancy is 75.63 years and population is 38.5 million people. Republic is an instance of qualitative data that results from observing the value of the quantitative variable government type. The life expectancy of 75.63 years is an instance of quantitative data that results from observing the value of the quantitative variable life expectancy.

## Determine the Level of Measurement of a Variable

### Definitions

A variable is at the ***nominal level of measurement*** if the values of the variable name, label, or categorize. In addition, the naming scheme does not allow for the values of the variable to be arranged in a ranked, or specific, order.

A variable is at the ***ordinal level of measurement*** if it has the properties of the nominal level of measurement and the naming scheme allows for the values of the variable to be arranged in a ranked, or specific, order.

A variable is at the ***interval level of measurement*** if it has the properties of the ordinal level of measurement and the differences in the values of the variable have meaning. A value of zero in the interval level of measurement does not mean the absence of the quantity. Arithmetic operations such as addition and subtraction can be performed on values of the variable.

**Example:** The years 2000, 1776, and 1492. Time did not begin in the year 0, so the year 0 is arbitrary instead of being a natural zero starting point representing “no time”.

A variable is at the ***ratio level of measurement*** if it has the properties of the interval level of measurement and the ratios of the values of the variable have meaning. A value of zero in the ratio level of measurement means the absence of the quantity. Arithmetic operations such as multiplication and division can be performed on the values of the variable.

**Example:** Prices of college textbooks (\$0 represents no cost, a \$100 book costs twice as much as a \$50 book)

**Example:** Distances (in km) traveled by cars (0 km represents no distance traveled, and 400 km is twice as far as 200 km).

<b><i>Levels of Measurement</i></b>		
<b><i>Ratio</i></b>	There is a natural zero starting point and ratios are meaningful	Distances
<b><i>Interval</i></b>	Differences are meaningful, but there is no natural zero starting point and ratios are meaningless	Body temperatures
<b><i>Ordinal</i></b>	Categories are ordered, but differences can't be found or are meaningless	Ranks of colleges
<b><i>Nominal</i></b>	Categories only. Data cannot be arranged in an ordering scheme	Eye colors

### Example

Body temperatures of 98.2°F and 98.6°F are examples of data at this interval level of measurement. Those values are ordered, and we can determine their difference of 0.4°F. However, there is no natural starting point. The value of 0°F might seem like a starting point, but it is arbitrary and does not represent the total absence of heat.



### ***Example***

*U.S. News* and *World Report* ranks colleges. Those ranks (first, second, third, and so on) determine an ordering. However, the differences between ranks are meaningless. For example, a difference of “second minus first” might suggest  $2 - 1 = 1$ , but this difference of 1 is meaningless because it is not an exact quantity that can be compared to other such differences. The difference between Harvard and Brown cannot be quantitatively compared to the difference between Yale and Johns Hopkins.

### ***Example***

Determine the level of measurement of the following variables considered in the study.

- a) Gender
- b) Temperature
- c) Number of days during the past week that a college student studied
- d) Letter grade
- e) Number of snack and soft drink vending machines in the school
- f) Whether or not the school has a closed campus policy during lunch
- g) Class rank (Freshman, Sophomore, Junior, Senior)
- h) Number of days per week a student eats school lunch

### **Solution**

- a) Gender is a variable measured at the ***nominal*** level because it only allows for categorization of male or female. Plus, it is not possible to rank gender classifications.
- b) Temperature is a variable measured at the ***interval*** level because differences in the value of the variable make sense.
- c) Number of days during the past week that a college student studied is measured at the ***ratio*** level, because the ratio of two values makes sense and a value of zero has meaning. For example, a student who studies four days studies twice as many days as a student who studies two days.
- d) Letter grade is a variable measured at the ***ordinal*** level because these grades can be arranged in order, but we can't determine differences between the grades. For example, we know that *A* is higher than *B* (so there is an ordering), but we cannot subtract *B* from *A* (so the difference cannot be found)
- e) Number of snack and soft drink vending machines in the school is a ***ratio*** measured
- f) Whether or not the school has a closed campus policy during lunch is a ***nominal*** measured
- g) Class rank (Freshman, Sophomore, Junior, Senior) is a ***ordinal*** measured
- h) Number of days per week a student eats school lunch is a ***ratio*** measured



## **Exercises**      **Section 1.1 – Introduction to the Practice Statistical**

1. Use common sense to determine whether the given event is **(a) impossible**; **(b) possible, but very unlikely**; **(c) possible and likely**.
  - a) Giants best the Denver Broncos in the Super Bowl by a score of 120 to 98.
  - b) While driving to his home in Connecticut, David was ticketed for driving 205 *mi/h* on a highway with a speed limit of 55 *mi/h*.
  - c) Thanksgiving Day will fall on a Monday next year.
  - d) When each of 25 statistics students turns on his or her TI-84 Plus calculator, all 25 calculators operate successfully.
  
2. Determine whether the underline value is a **parameter** or a **statistic**.
  - a) Following the 2010 national midterm election, 18% of the governors of the 50 United States were female.
  - b) The average score for a class of 28 students taking a calculus midterm exam was 72%.
  - c) In a national survey of 1300 high school students (grades 9 to 12), 32% of respondents reported that someone has bullied them at school.
  - d) In a national survey on substance abuse, 10.0% of respondents aged 12 to 17 reported using illicit drugs within the past month.
  - e) Ty Cobb is one of major league baseball's greatest hitters of all time, with a career batting average of 0.366.
  - f) Only 12 men have walked on the moon. The average age of these men at the time of their moonwalks was 39 years, 11 months, 15 days.
  - g) A study of 6076 adults in public rest rooms (in Atlanta, Chicago, New York City, and San Francisco) found that 23% did not wash their hands before exiting.
  - h) Interviews of 100 adults 18 years of age or older, conducted nationwide, found that 44% could state the minimum age required for the office of U.S. president.
  
3. Classify the variable as **qualitative** or **quantitative**
  - a) Nation of origin
  - b) Number of siblings
  - c) Grams of carbohydrates in a doughnut
  - d) Number on a football player's jersey
  - e) Number of unpopped kernels in a bag of ACT microwave popcorn
  - f) Assessed value of a house
  - g) Phone number
  - h) Student ID number.
  - i) Favorite film
  - j) Population of country of origin
  - k) Gallons of water in a swimming pool
  - l) Model of car driven
  - m) Distance in miles to nearest school
  - n) Time in hours that a light bulb lasts

- o) Number of students at a high school
4. Determine whether the quantitative variable is **discrete** or **continuous**
- Goals scored in a season by a soccer player
  - Volume of water lost each day through a leaky faucet
  - Length (in minutes) of a country song
  - Number of Sequoia trees in a randomly selected acre of Yosemite National Park
  - Temperature on a randomly selected day in Memphis, Tennessee
  - Internet connection speed in Kilobytes per second
  - Points scored in an NCAA basketball game
  - Air pressure in pounds per square inch in an automobile tire
5. Determine the level of measurement of each variable
- Nation of origin
  - Movie ratings of one star through five stars
  - Volume of water used by a household in a day
  - Year of birth of college students
  - Highest degree conferred (high school, bachelor's, and so on)
  - Eye color
  - Assesses value of a house
  - Time of day measured in military time
6. The Gallup Organization contacts 1026 teenagers who are 13 to 17 years of age and live in the United States and asks whether or not they had been prescribed medications for any mental disorders, such as depression or anxiety. Identify the population and sample.
7. A quality-control manager randomly selects 50 bottles of Coca-Cola that were filled on October 15 to assess the calibration of the filling machine. Identify the population and sample.

### Exercise 8-9

*Nicotine Amounts from Menthol and King-Size Cigarettes*

<b>x</b>	1.1	0.8	1.0	0.9	0.8
<b>y</b>	1.1	1.7	1.7	1.1	1.1

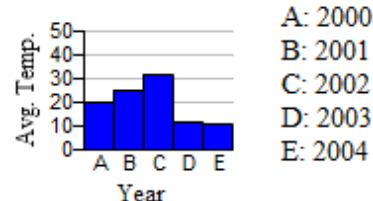
The  $x$ -values are nicotine amounts (in  $mg$ ) in different 100  $mm$  filtered, non-light menthol cigarettes; the  $y$ -values are nicotine amounts (in  $mg$ ) in different king-size non-filtered, non-menthol, and non-light cigarettes.

8. Each  $x$  value associated with the corresponding  $y$  value in some meaningful way? If the  $x$  and  $y$  values are not matched, does it make sense to use the difference between each  $x$  value and the  $y$  value that is the same column?
9. The Federal Trade Commission obtained the measured amounts of nicotine in the table. Is the source of the data likely to be unbiased?
- Note that the table lists measured nicotine amounts from two different types of cigarette. Given these data, what issue can be addressed by conducting a statistical analysis of the values?

10. One of Gregor Mendel's famous hybridization experiments with peas yielding 580 off spring with 152 of those peas (or 26%) having yellow pods. According to Mendel's theory, 25% of the off spring peas should have yellow pods. Do the results of the experiment differ from Mendel's claimed rate of 25% by an amount that is statistically significant?
11. In a Gallup poll of 1038 randomly selected adults, 85% said that secondhand smoke is somewhat harmful or very harmful, but a representative of the tobacco industry claims that only 50% of adults believe that secondhand smoke is somewhat harmful or very harmful. Is there statistically significant evidence against the representative's claim? Why or why not?
12. Determine whether the given value is parametric or a statistic
  - a) One of greatest baseball hitters of all time has a career batting average of 0.366
  - b) A sample of employees is selected and it is found that 50% own a vehicle
  - c) A survey of 42 out of hundreds in a dining hall showed that 17 enjoyed their meal
13. Suppose a survey of 568 women in the U.S. found that more than 61% are the primary investor in their household.
  - a) Describe the survey represents the descriptive branch of statistic
  - b) Make an inference based on the results of the survey
14. In the recent study, volunteers who had 8 hours of sleep were three times more likely to answer questions correctly on a math test than were sleep-deprived participants.
  - a) Identify the sample used in the study
  - b) What is the sample's population
  - c) Which part of the study represents the descriptive branch of statistics
  - d) Make an inference based on the results of the study
15. Determine whether the data set is a population or a sample. Explain your reasoning  
The salary if each baseball player in a league
16. In a poll, 1,004 adults in a country were asked whether they favor or oppose the use of "federal tax dollars to fund medical research using stem cells obtained from human embryos." Among the responders, 48% said that they were in favor. Describe the statistical study
  - a) What is the population?
  - b) Identify the sample
17. A study shows that the obesity rate among boys ages 2 to 19 has increased over the past several years
  - a) Make an inference based on the results of this study?
  - b) What is wrong with this type of reasoning
18. The newspaper USA Today published a health survey, and some readers completed the survey and returned it. Identify the (a) sample and (b) population, also determine whether the sample likely to be representative of the population.

19. A Gallup poll of 1012 randomly surveyed adults found that 9% of them said cloning of humans should be allowed. Identify the (a) sample and (b) population, also determine whether the sample likely to be representative of the population.
20. Some people responded to this request: “Dial 1-900-PRO-LIFE to participate in a telephone poll on abortion. (\$1.95 per minute. Average call: 2 minutes. You must be 18 years old.)” Identify the (a) sample and (b) population, also determine whether the sample likely to be representative of the population
21. In the Born Loser cartoon strip by Art Sansom, Brutus expresses joy over an increase in temperature from 1° to 2°. When asked what is so good about 2°, he answers that “it’s twice as warm as this morning.” explain why Brutus is wrong yet again.
22. A group of students develops a scale for rating the quality of cafeteria food, with 0 representing “neutral: not good and not bad.” Bad meals are given negative numbers and good meals are given positive numbers, with the magnitude of the number corresponding to the severity of badness or goodness. The first three meals are rated as 2, 4, and –5. What is the level of measurement for such rating? Explain your choice.
23. Suppose that a study based on a sample from a targeted population shows that people who own a fax machine have more money than people who do not
  - a) Make an inference based on the results of this study?
  - b) What might this inference incorrectly imply?
24. Determine whether the statement is true or false, rewrite it as a true statement
  - a) Data at the ordinal level are quantitative only
  - b) More types of calculations can be performed with data at the nominal level than with data at the interval level
25. The region of a country with the highest per capita income for the past six years is shown below  
 Northeast    Southern    Southwest    Southeast    Northern    Western
  - a) Determine whether the data are qualitative or quantitative and identify the data set’s level of measurement
  - b) What is the data set’s level of measurement?
26. The region of a country with the six highest level of food production last year are shown below  
 1. Eastern    2. Southwest    3. Western    4. Southeast    5. Northwest    6. Southern
  - a) Determine whether the data are qualitative or quantitative and identify the data set’s level of measurement
  - b) What is the data set’s level of measurement?
27. The region of a country with the six highest level of food production last year are shown below  
 22.8    26.4    24.1    22.2    21.6    21.1    25.8    21.5    24.6
  - a) Determine whether the data are qualitative or quantitative and identify the data set’s level of measurement
  - b) What is the data set’s level of measurement?

28. The graph shows the average temperature in an arctic city, in degree Fahrenheit, for certain years. Identify the level of measurement of the data listed on the horizontal axis in the graph



29. Identify the level of measurement of the data:

- a) Temperature
- b) Age
- c) Family history of illness
- d) Pain level (scale of 0 to 10)

30. A study was conducted in which 20,211 18-years old male military were given an exam to measure IQ. In addition, the recruits were asked to disclose their smoking status. An individual was considered a smoker if he smoked at least once cigarette per day. The goal of the study was to determine whether adolescents aged 18 to 21 who smoked have a lower IQ than nonsmokers. It was found that the average IQ of the smokers was 94, while he average IQ of the nonsmokers was 101. The researchers concluded that lower IQ individuals are more likely to choose to smoke, not that smoking makes people less intelligent.

- a) What is the research objective?
- b) What is the population being studied? What is the sample?
- c) What are the descriptive statistics?
- d) What are the conclusions of the study?

31. Determine whether the variable is qualitative, continuous, or discrete. The following represent information on smart phones.

<i>Model</i>	<i>Weight (oz.)</i>	<i>Service Provider</i>	<i>Depth (in)</i>
Motorola Droid X	5.47	Verizon	0.39
Motorola Droid 2	5.96	Verizon	0.53
Apple iPhone 4	4.8	ATT	0.37
Samsung Epic 4G	5.5	Sprint	0.6
Samsung Captivate	4.5	ATT	0.39

## Section 1.2 – Observational Studies versus Designed Experiments

### Distinguish between an Observational Study and an Experiment

#### Basics of Collecting Data

Statistical methods are driven by the data that we collect. We typically obtain data from two distinct sources: *observational studies* and *experiment*.

#### Example

Researchers Joachim Schüz and associates wanted “to investigate cancer risk among Danish cellular phone users who were followed for up to 21 years.” To do so, they kept track of 420,095 people whose first cellular telephone subscription was between 1982 and 1995. In 2002, they recorded the number of people out of the 420,095 people who had a brain tumor and compared the rate of brain tumors in this group to the rate of brain tumors in the general population.

#### Example

They found no significant difference in the rate of brain tumors between the two groups. The researchers concluded “cellular telephone was not associated with increased risk for brain tumors.” (Source: Joachim Schüz et al. “Cellular Telephone Use and Cancer Risk: Update of a Nationwide Danish Cohort,” *Journal of the National Cancer Institute* 98(23): 1707-1713, 2006)

Researchers Joseph L. Roti and associates examined “whether chronic exposure to radio frequency (RF) radiation at two common cell phone signals—835.62 megahertz, a frequency used by analogue cell phones, and 847.74 megahertz, a frequency used by digital cell phones—caused brain tumors in rats. The rats in group 1 were exposed to the analogue cell phone frequency; the rats in group 2 were exposed to the digital frequency; the rats in group 3 served as controls and received no radiation. The exposure was done for 4 hours a day, 5 days a week for 2 years. The rats in all three groups were treated the same, except for the RF exposure.

After 505 days of exposure, the researchers reported the following after analyzing the data. “We found no statistically significant increases in any tumor type, including brain, liver, lung or kidney, compared to the control group.” (Source: M. La Regina, E. Moros, W. Pickard, W. Straube, J. L. Roti Roti. “The Effect of Chronic Exposure to 835.62 MHz FMCW or 847.7 MHz CDMA on the incidence of Spontaneous Tumors in Rats.” Bioelectromagnetic Society Conference, June 25, 2002.)

In both studies, the goal of the research was to determine if radio frequencies from cell phones increase the risk of contracting brain tumors. Whether or not brain cancer was contracted is the **response variable**. The level of cell phone usage is the **explanatory variable**.

In research, we wish to determine how varying the amount of an **explanatory variable** affects the value of a **response variable**.

#### Definitions

An **observational study** measures the value of the response variable without attempting to influence the value of either the response or explanatory variables. That is, in an observational study, the researcher observes the behavior of the individuals in the study without trying to influence the outcome of the study.

**Example:** The Literary Digest poll in which respondents were asked who they would vote for in the presidential election is an observational study. The subjects were asked for their choices, but they were not given any type of treatment.

If a researcher assigns the individuals in a study to a certain group, intentionally changes the value of the explanatory variable, and then records the value of the response variable for each group, the researcher is conducting a **designed experiment**.

**Experiment** apply some treatment and then observe its effects on the subjects; (subjects in experiments are called experimental units)

**Example:** In the largest public health experiment ever conducted, 200,745 children were given a treatment consisting of the Salk vaccine, while 201,229 other children were given a placebo. The Salk vaccine injections constitute a treatment that modified the subjects, so this is an example of an experiment.

### **Example**

Researchers wanted to determine the long-term benefits of the influenza vaccine on seniors aged 65 years and older. The researchers looked at records of over 36,000 seniors for 10 years. The seniors were divided into two groups. Group 1 were seniors who chose to get a flu vaccination shot, and group 2 were seniors who chose not to get a flu vaccination shot. After observing the seniors for 10 years, it was determined that seniors who get flu shots are 27% less likely to be hospitalized for pneumonia or influenza and 48% less likely to die from pneumonia or influenza. (Source: Kristin L. Nichol, MD, MPH, MBA, James D. Nordin, MD, MPH, David B. Nelson, PhD, John P. Mullooly, PhD, Eelko Hak, PhD. “Effectiveness of Influenza Vaccine in the Community-Dwelling Elderly,” New England Journal of Medicine 357:1373–1381, 2007)

Based on the results of this study, would you recommend that all seniors go out and get a flu shot? The study may have flaws! Namely, *confounding*.

### **Definitions**

**Confounding** in a study occurs when the effects of two or more explanatory variables are not separated. Therefore, any relation that may exist between an explanatory variable and the response variable may be due to some other variable or variables not accounted for in the study.

A **lurking variable** is an explanatory variable that was not considered in a study, but that affect the value of the response variable in the study. In addition, lurking variables are typically related to any explanatory variables considered in the study.

- ✓ Some lurking variables in the influenza study: age, health status, or mobility of the senior
- ✓ Even after accounting for potential lurking variables, the authors of the study concluded that getting an influenza shot is *associated* with a lower risk of being hospitalized or dying from influenza.
- ✓ Observational studies do not allow a researcher to claim causation, only association.



## Explain the Various Types of Observational Studies

**Cross-sectional Studies** Observational studies that collect information about individuals at a specific point in time, or over a very short period of time.

**Case-control Studies** These studies are *retrospective*, meaning that they require individuals to look back in time or require the researcher to look at existing records. In case-control studies, individuals who have certain characteristics are matched with those that do not.

**Cohort Studies** A cohort study first identifies a group of individuals to participate in the study (the cohort). The cohort is then observed over a long period of time. Over this time period, characteristics about the individuals are recorded. Because the data is collected over time, cohort studies are *prospective*.

### Example

Determine whether each of the following studies depict an observational study or an experiment. If the researchers conducted an observational study, determine the type of the observational study.

- a) Researchers wanted to assess the long-term psychological effects on children evacuated during World War II. They obtained a sample of 169 former evacuees and a control group of 43 people who were children during the war but were not evacuated. The subjects' mental states were evaluated using questionnaires. It was determined that the psychological well being of the individuals was adversely affected by evacuation. (Source: Foster D, Davies S, and Steele H (2003) The evacuation of British children during World War II: a preliminary investigation into the long-term psychological effects. *Aging & Mental Health* (7)5.)

**Observational study; Case-control**

- b) Xylitol has proven effective in preventing dental caries (cavities) when included in food or gum. A total of 75 Peruvian children were given milk with and without xylitol and were asked to evaluate the taste of each. Overall, the children preferred the milk flavored with xylitol. (Source: Castillo JL, et al (2005) Children's acceptance of milk with xylitol or sorbitol for dental caries prevention. *BMC Oral Health* (5)6.)

**Designed experiment**

- c) A total of 974 homeless women in the Los Angeles area were surveyed to determine their level of satisfaction with the healthcare provided by shelter clinics versus the healthcare provided by government clinics. The women reported greater quality satisfaction with the shelter and outreach clinics compared to the government clinics. (Source: Swanson KA, Andersen R, Gelberg L (2003) Patient satisfaction for homeless women. *Journal of Women's Health* (12)7.)

**Observational study; Cross-sectional**

- d) The Cancer Prevention Study II (CPS-II) is funded and conducted by the American Cancer Society. Its goal is to examine the relationship among environmental and lifestyle factors on cancer cases by tracking approximately 1.2 million men and women. Study participants completed an initial study questionnaire in 1982 providing information on a range of lifestyle factors such as diet, alcohol and tobacco use, occupation, medical history, and family cancer history. These data have been examined extensively in relation to cancer mortality. Vital status of study participants is updated biennially.

Cause of death has been documented for over 98% of all deaths that have occurred. Mortality follow-up of the CPS-II participants is complete through 2002 and is expected to continue for many years.

(Source: American Cancer Society)

***Observational study; cohort***

### ***Definition***

A ***census*** is a list of all individuals in a population along with certain characteristics of each individual.

## **Exercises**    **Section 1.2 – Observational Studies vs Designed Experiments**

1. Researchers wanted to know if there is a link between proximity to high-tension wires and the rate of leukemia in children. To conduct the study, researchers compared the rate of leukemia for children who lived within  $\frac{1}{2}$  mile of high-tension wires to the rate of leukemia for children who did not live within  $\frac{1}{2}$  mile of high-tension wires. Determine whether the study depicts an observational study or an experiment.
2. Rats with cancer are divided into two groups. One group receives 5 milligrams (mg) of a medication that is thought to fight cancer, and the other receives 10 mg. After 2 years, the spread of the cancer is measured. Determine whether the study depicts an observational study or an experiment.
3. Seventh-grade students are randomly divided into two groups. One group is taught math using traditional techniques; the other is taught math using a reform method. After 1 year, each group is given an achievement test to compare proficiency. Determine whether the study depicts an observational study or an experiment.
4. A poll is conducted in which 500 people are asked whom they plan to vote for in the upcoming election. Determine whether the study depicts an observational study or an experiment.
5. A survey is conducted asking 400 people. “Do you prefer Coke or Pepsi?” Determine whether the study depicts an observational study or an experiment.
6. A Gallup poll surveyed 1018 adults by telephone, and 22% of them reported that they smoked cigarettes within the past year. Determine whether the description corresponds to an observation study or an experiment.
7. In a morally and criminally wrong study, 399 black men with syphilis were not given a treatment that could have cured them. The intent was to learn about the effects of syphilis on black men. The subjects were initially treated with small amounts of bismuth, neoarsphenamine, and mercury, but those treatments were replaced with aspirin. Determine whether the description corresponds to an observation study or an experiment.
8. While shopping, 200 people are asked to perform a taste test in which they drink from two randomly placed, unmarked cups. They are then asked which drink they prefer. Determine whether the description corresponds to an observation study or an experiment.
9. Conservation agents netted 250 large-mouth bass in a lake and determined how many were carrying parasites. Determine whether the description corresponds to an observation study or an experiment.
10. Researchers wanted to determine if there was an association between the level of happiness of an individual and their risk of heart disease. The researchers studied 1739 people over the course of 10 years. During this 10-year period, they interviewed the individuals and asked questions about their daily lives and the hassles they face. In addition, hypothetical scenarios were presented to determine how each individual would handle the situation. These interviews were videotaped and studied to

assess the emotions of the individuals. The researchers also determined which individuals in the study experienced any type of heart disease over the 10-year period. After their analysis, the researchers concluded that the happy individuals were less likely to experience heart disease.

- a) What type of observational study is this? Explain.
- b) What is the response variable?
- c) What is the explanatory variable?
- d) In the report, the researchers stated that “the research team also hasn’t ruled out that a common factor like genetics could be causing both the emotions and the heart disease.” Use the language introduced on this section to explain what this sentence means.

11. Researchers wanted to determine if there was an association between daily coffee consumption and the occurrence of skin cancer. The researchers looked at 93,676 women enrolled in the Women’s Health Initiative Observation Study and asked them to report their coffee-drinking habits. The researchers also determined which of the women had nonmelanoma skin cancer. After their analysis, the researchers concluded that consumption of six or more cups of caffeinated coffee per day was associated with a reduction in nonmelanoma skin cancer

- a) What type of observational study is this? Explain.
- b) What is the response variable?
- c) What is the explanatory variable?
- d) In their report, the researchers stated that “After adjusting for various demographic and lifestyle variables, daily consumption of six or more cups was associated with a 30% reduced prevalence of nonmelanoma skin cancer.” Why was it important to adjust for these variables?

12. Researcher Penny Gordon-Larson and her associate wanted to determine whether young couples who marry or cohabitate are more likely to gain weight than those who stay single. The researchers followed 8000 men and women for 7 years as they matured from teens to young adults. When the study began, none of the participants were married or living with a romantic partner. By the end of the study, 14% of the participants were married and 16% were living with a romantic partner. The researchers found that married or cohabiting women gained, on average, 9 pounds more than single.

- a) Why is this an observation study? What type of observational study is this?
- b) What is the response variable in the study?
- c) What is the explanatory variable?
- d) Identify some potential lurking variables in this study.
- e) Can we conclude that getting married or cohabiting causes one to gain weight? Explain.

## Section 1.3 – Sampling Methods

### Obtain a Simple Random Sample

#### *Definitions*

**Random sampling** is the process of using chance to select individuals from a population to be included in the sample.

If convenience is used to obtain a sample, the results of the survey are meaningless.

A selection so that each individual member has an equal chance of being selected



A sample of size  $n$  from a population of size  $N$  is obtained through **simple random sampling** if every possible sample of size  $n$  has an equally likely chance of occurring. The sample is then called a **simple random sample**.

**Probability Sample** selecting members from a population in such a way that each member of the population has a known (but not necessarily the same) chance of being selected

#### *Example*

Suppose a study group consists of 5 students: Bob, Patricia, Mike, Jan, and Maria

2 of the students must go to the board to demonstrate a homework problem. List all possible samples of size 2 (without replacement).

#### *Solution*

- Bob, Patricia
- Bob, Mike
- Bob, Jan
- Bob, Maria
- Patricia, Mike
- Patricia, Jan
- Patricia, Maria
- Mike, Jan
- Mike, Maria
- Jan, Maria

#### *Example*

Sophia has four tickets to a concert. Six of her friends, Yolanda, Michael, Kevin, Marissa, Annie, and Katie, have all expressed an interest in going to the concert. Sophia decides to randomly select three of her six friends to attend the concert.

- a) List all possible samples of size  $n = 3$  from the population of size  $N = 6$ . Once an individual is chosen, he or she cannot be chosen again.
- b) Comment on the likelihood of the sample containing Michael, Kevin, and Marissa.

#### *Solution*

- a) The possible samples of size 3 are:

Yolanda, Michael, Kevin	Yolanda, Michael, Marissa	Yolanda, Michael, Annie	Yolanda, Michael, Katie
Yolanda, Kevin, Marissa	Yolanda, Kevin, Annie	Yolanda, Kevin, Katie	Yolanda, Kevin, Annie
Yolanda, Marissa, Katie	Yolanda, Annie, Katie	Michael, Kevin, Marissa	Michael, Kevin, Annie
Michael, Kevin, Katie	Michael, Marissa, Annie	Michael, Marissa, Katie	Michael, Annie, Katie
Kevin, Marissa, Annie	Kevin, Marissa, Katie	Kevin, Annie, Katie	Marissa, Annie, Katie

There are 20 possible sample size 3 from population of 6.

- b) Only 1 of the 20 possible samples contains Michael, Kevin, and Marissa, so there is a 1 in 20 chance that the simple random sample will contain these three. In fact, all the samples of size 3 have a 1 in 20 chance of occurring.

## Steps for Obtaining a Simple Random Sample

1. Obtain a frame that lists all the individuals in the population of interest. Number the individuals in the frame 1 –  $N$ .
2. Use a random number table, graphing calculator, or statistical software to randomly generate  $n$  numbers where  $n$  is the desired sample size.

### Example

The 112th Congress of the United States had 435 members in the House of Representatives. Explain how to conduct a simple random sample of 5 members to attend a Presidential luncheon. Then obtain the sample.

### Solution

**Step 1** Put the members in alphabetical order. Number the members from 1 - 435.

**Step 2** Randomly select five numbers using a random number generator. First, set the seed. The seed is an initial point for the generator to start creating random numbers—like selecting the initial point in the table of random numbers. The seed can be any nonzero number. Then generate the random numbers. Match the generated random numbers to the corresponding Representatives.

### Example

The accounting firm of Sense and Associates has grown. To make sure their clients are still satisfied with the services that are receiving, the company decides to send a survey out to a simple random sample of 5 of its 30 clients. Find a simple random sample of five clients.

### Solution

**Step 1:** We can create a client list and number them from 01 to 10.

**Step 2:** A table of random numbers can be used to select the individuals to be in the sample.



The following numbers are: 11, 4, 20, 29, 11, 27

We ignore the second 11 because we are sampling without replacement. The clients corresponding to these numbers are the clients to be surveyed.

## 2. Obtain a Stratified Sample

### *Definition*

A **stratified sample** is one obtained by separating the population into homogeneous, nonoverlapping groups called strata, and then obtaining a simple random sample from each stratum. The individuals within each stratum should be homogeneous (or similar) in some way.

*Stratified Sampling* subdivide the population into at least two different subgroups that share the same characteristics, then draw a sample from each subgroup (or stratum)



### *Example*

In 2008, the United States Senate had 47 Republicans, 51 Democrats, and 2 Independents. The president wants to have a luncheon with 4 Republicans, 4 Democrats and 1 Other. Obtain a stratified sample in order to select members who will attend the luncheon.

To obtain the stratified sample, conduct a simple random sample within each group. That is, obtain a simple random sample of 4 Republicans (from the 47), a simple random sample of 4 Democrats (from the 51), and a simple random sample of 1 Other from the 100. Be sure to use a different seed for each stratum.

### *Example*

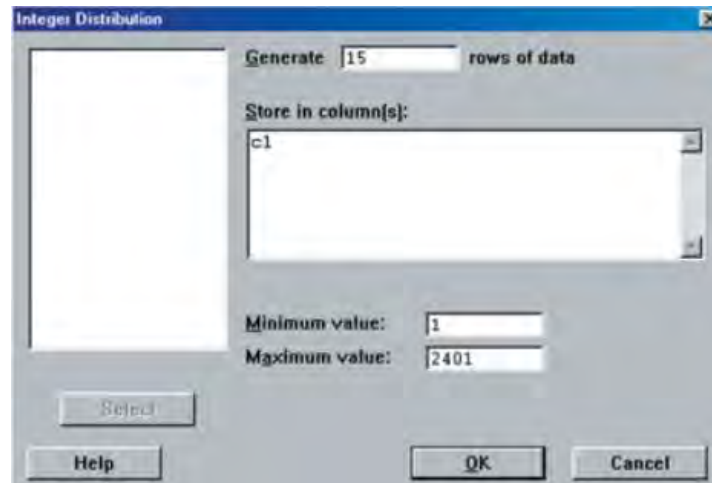
The president of DePaul University wants to conduct a survey to determine the community's opinion regarding campus safety. The president divides the DePaul community into three groups: resident students,, nonresident (commuting) students, and staff (including faculty) so that he can obtain a stratified sample, Suppose there are 6204 resident students, 13,304 nonresident students, and 2401 staff, for a total of 21,909 individuals in the population. The president wants to obtain a sample of size 100, with the number of individuals selected from each stratum weighted by the population size. So resident students



make up  $\frac{6204}{21,909} = 28\%$ , nonresident students account for 61% and staff constitute 11% of the sample. A sample of size 100 requires a stratified sample of  $0.28(100) = 28$  resident students,  $0.61(100) = 61$  nonresident, and  $0.11(100) = 11$  staff.

### **Solution**

Using MINITAB, with the seed set to 4032 and the values, we can obtain the following sample of staff: 240, 230, 847, 190, 2096, 705, 2320, 323, 701, 471, 744



Repeat this proceed for the resident and nonresident students using a different seed.

## **3. Obtain a Systematic Sample**

### ***Definition***

A **systematic sample** is obtained by selecting every  $k^{\text{th}}$  individual from the population. The first individual selected is a random number between 1 and  $k$ .

*Systematic Sampling* Select some starting point and then select every  $k$ th element in the population



### **Steps in Systematic Sampling**

**Step 1:** Determine the population size,  $N$ .

**Step 2:** Determine the sample size desired,  $n$ .

**Step 3:** Compute  $N/n$  and round down to the nearest integer. This value is  $k$ .

**Step 4:** Randomly select a number between 1 and  $k$ . Call this number  $p$ .

**Step 5:** The sample will consist of the following individuals:

$$p, p + k, p + 2k, \dots, p + (n - 1)k$$

### Example

A quality control engineer wants to obtain a systematic sample of 25 bottles coming off a filling machine to verify the machine is working properly. Design a sampling technique that can be used to obtain a sample of 25 bottles.

### Solution

If we choose a number between 1 and 7, say 5. Then

5,  $5 + 7 = 12$ , 19, 26, ..., 173

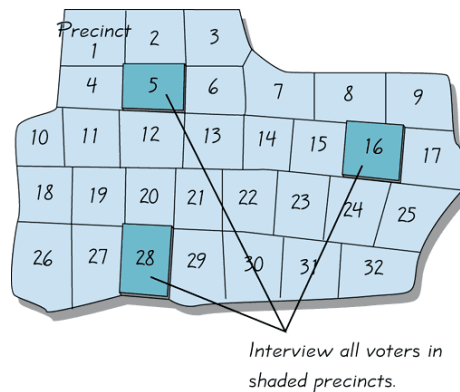
$$p, p + k, p + 2k, \dots, p + (n - 1)k$$

## 4. Obtain a Cluster Sample

### Definition

A **cluster sample** is obtained by selecting all individuals within a randomly selected collection or group of individuals.

**Cluster Sampling** divide the population area into sections (or clusters); randomly select some of those clusters; choose all members from selected clusters



**Multistage Sampling** Collect data by using some combination of the basic sampling methods

In a multistage sample design, pollsters select a sample in different stages, and each stage might use different methods of sampling.

### Example

A sociologist wants to gather data regarding household income within the city of Boston. Obtain a sample using cluster sampling

### Solution

Suppose there are 10,493 city blocks in Boston. First, the sociologist must number the blocks from 1 to 10,493. Suppose the sociologist has enough time and money to survey 20 clusters (city blocks). The sociologist should obtain a simple random sample of 20 numbers between 1 and 10,493 and survey all households from the clusters selected. Cluster sampling is a good choice in this example because it reduces the travel time to households that is likely to occur with both simple random sampling and

stratified sampling. In addition, there is no need to obtain a frame of all the households with cluster sampling. The only frame needed is one that provides information regarding city blocks

### Example

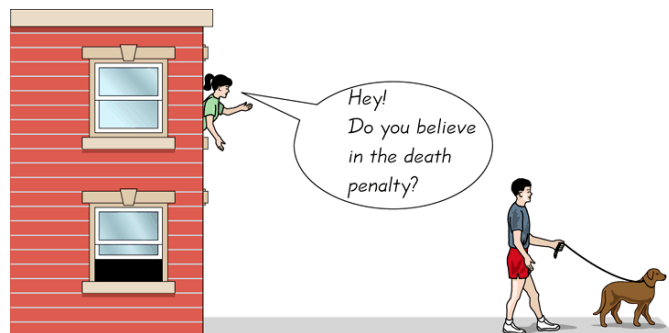
The U.S. government's unemployment statistics are based on surveyed households. It is impractical to personally visit each member of a simple random sample, because individual households would be spread all over the country. Instead, the U.S. Census Bureau and the Bureau of Labor Statistics combine to conduct a survey called the Current Population Survey. This survey obtains data describing such factors as unemployment rates, college enrollments, and weekly earning amounts. The survey incorporates a multistage sample design, roughly following these steps:

1. The surveys partition the entire United States into 2007 different regions called *primary sampling units* (PSU). The primary sampling units are metropolitan areas, large counties, or groups of smaller counties.
2. The surveyors select a sample of primary sampling units in each of the 50 states. For the Current Population Survey, 792 of the primary sampling units are used. ( All of the 432 primary sampling units with the largest populations are used, and 360 primary sampling units are randomly selected from the other 1575.)
3. The surveyors partition each of the 792 selected primary sampling units into blocks, and they then use stratified sampling to select a sample of blocks.
4. In each selected block, surveyors identify clusters of households that are close to each other. They randomly select clusters, and they interview all households in the selected clusters.

This multistage sample design includes random, stratified, and cluster sampling at different stages. The end result is a complicated sampling design, but it is much more practical and less expensive than using a simpler design, such as using a simple random sample.

### Definition

A **convenience sample** is one in which the individuals in the sample are easily obtained. *Convenience Sampling* use results that are easy to get



Any studies that use this type of sampling generally have results that are suspect. Results should be looked upon with extreme skepticism.

## Example

In practice, most large-scale surveys obtain samples using a combination of the techniques just presented. As an example of multistage sampling, consider Nielsen Media Research. Nielsen randomly selects households and monitors the television programs these households are watching through a People Meter. The meter is an electronic box placed on each TV within the household. The People Meter measures what program is being watched and who is watching it. Nielsen selects the households with the use of a two-stage sampling process.

## Solution

**Stage 1** Using U.S. Census data, Nielsen divides the country into geographic areas (strata). The strata are typically city blocks in urban areas and geographic regions in rural areas. About 6000 strata are randomly selected.

**Stage 2** Nielsen sends representatives to the selected strata and lists the households within the strata. The households are then randomly selected through a simple random sample.

Nielsen sells the information obtained to television stations and companies. These results are used to help determine prices for commercials.

As another example of multistage sampling, consider the sample used by the Census Bureau for the Current Population Survey. This survey requires five sample stages of sampling:

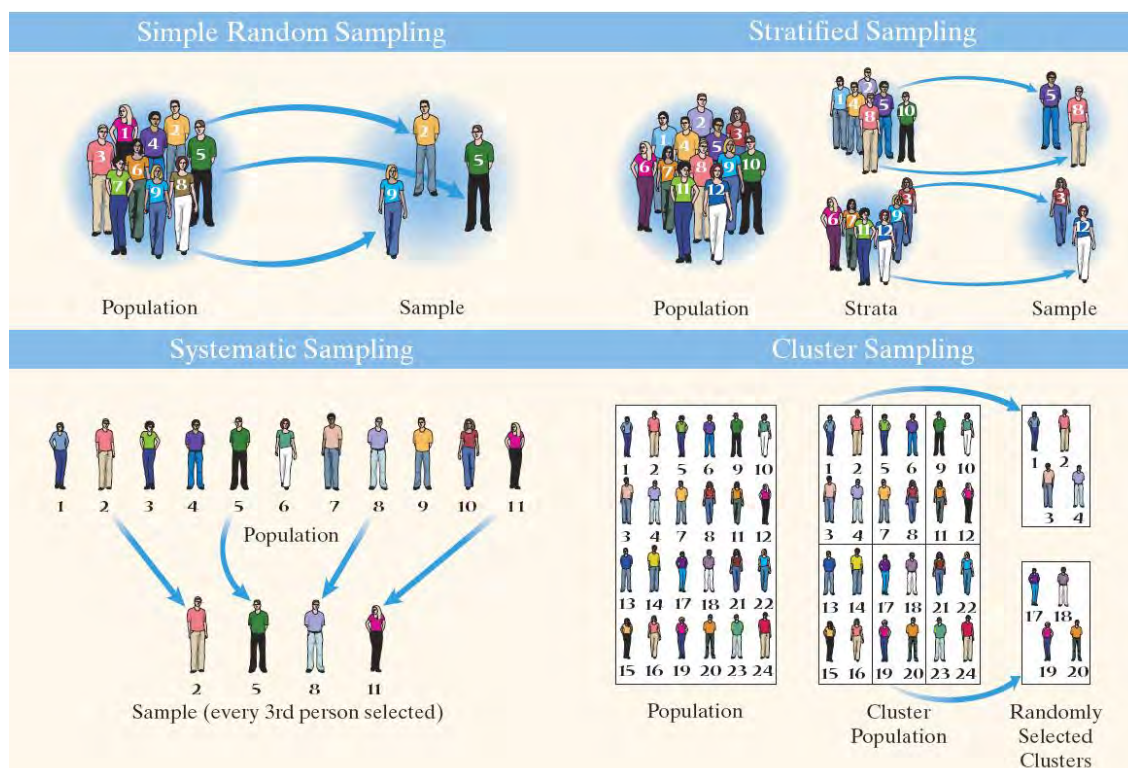
**Stage 1:** Stratified sample

**Stage 2:** Cluster sample

**Stage 3:** Stratified sample

**Stage 4:** Cluster sample

**Stage 5:** Systematic sample



## ***Exercises***     **Section 1.3 – Sampling Methods**

1. A student of the author collected measurements of arm lengths from her family members. Identify what type is used: random, systematic, convenience, stratified, or cluster.
2. On the day of the last presidential election, ABC News organized an exit poll in which specific polling stations were randomly selected and all voters were surveyed as they left the premises. Identify what type is used: random, systematic, convenience, stratified, or cluster.
3. The author was an observer at a town of Poughkeepsic Police sobriety checkpoint at which every fifth driver was stopped and interviewed. (He witnessed the arrest of a former student.) Identify what type is used: random, systematic, convenience, stratified, or cluster.
4. You observed professional wine taster working at the Consumer's Union testing facility in NY. Assume that a taste test involves three different wines randomly selected from each of five different wineries. Identify what type is used: random, systematic, convenience, stratified, or cluster.
5. The U.S. Department of Corrections collects data about returning prisoners by randomly selecting five federal prisons and surveying all of the prisoners in each of the prisons. Identify what type is used: random, systematic, convenience, stratified, or cluster.
6. You instructor surveyed all of his students to obtain sample consisting of the number of credit cards students possess. Identify what type is used: random, systematic, convenience, stratified, or cluster.
7. In a study of college programs, 820 students are randomly selected from those majoring in communications, 1463 students are randomly selected from those majoring in business, and 760 students are randomly selected from those majoring in history. Identify what type is used: random, systematic, convenience, stratified, or cluster.
8. Pharmacists typically fill prescriptions by scooping a sample of pills from a larger batch that is in stock. A pharmacist thoroughly mixes a large batch of Lipitor pills, then selects 30 of them. Does this sampling plan result in a random sample? Simple random sample? Explain.
9. A quality control engineer selects every 10,000<sup>th</sup> M&M plain candy that is produced. Does this sampling plan result in a random sample? Simple random sample? Explain.
10. NBC News polled reactions to the last presidential election by surveying adults who were approached by a reporter at a location in N.Y. City. Does this sampling plan result in a random sample? Simple random sample? Explain.
11. A classroom consists of 36 students seated in six different rows, with six students in each row. The instructor rolls a die to determine a row, then rolls the die again to select a particular student in the row. This process is repeated until a sample of 6 students is obtained. Does this sampling plan result in a random sample? Simple random sample? Explain.

12. A computer company employs 100 software engineers and 100 hardware engineers. The personnel manager randomly selects 20 of the software engineers and 20 of the hardware engineers and questions them about career opportunities within the company. Does the sampling plan result in a random sample? Simple random sample? Explain.
13. A polling company obtains an alphabetical list of names of voters in a precinct. They select every 20<sup>th</sup> person from the list until a sample of 100 is obtained. They then call these 100 people. Does the sampling plan result in a random sample? Simple random sample? Explain.
14. What is an inherent zero? Describe three examples of data sets that have inherent zeros and three that do not.
15. What is the different between a random sample and a simple random sample?
16. Determine whether the statement is true or false. If false, rewrite it as a true statement
  - a) In a randomized block design, subjects with similar characteristics are divided into blocks, and then, within each block, randomly assigned to treatment groups.
  - b) Using a systematic sample guarantees that members of each group within a population will be sampled.
  - c) The method for selected a stratified sample is to order a population in some way and then select members of the population at regular intervals.
17. Which method of data collection should be used to collect data for the following study
  - a) A study of the health of 148 kidney transplant patients at a hospital.
  - b) A study of the effect on the taste of a snack food made with a sugar substitute
  - c) A study of how fast a virus would spread in a herd of cattle.
18. A pharmaceutical company wants to test the effectiveness of a new allergy drug. The company identifies 250 females 30-35 years old who suffer from severe allergies. The subjects are randomly assigned into two groups. One group is given the new allergy drug and the other is given a placebo that looks exactly like the new allergy drug. After six months, the subjects' symptoms are studied and compared
  - a) Identify the experimental units and treatment used in this experiment.
  - b) Identify a potential problem with the experiment design being used and suggest a way to improve it.
  - c) How could this experiment be designed to be a double-blind?
19. What type of sampling is used: random, stratified, convenience, cluster, systematic, in the following?
  - a) To estimate the percentage of defects in a recent manufacturing batch, a quality-control manager at Intel selects every 8<sup>th</sup> chip that comes off the assembly line starting with the 3<sup>rd</sup> until she obtains a sample of 140 chips.
  - b) To determine the prevalence of human growth hormone (HGH) use among high school varsity baseball players, the State Athletic Commission randomly selects 50 high schools. All members of the selected high schools' varsity baseball teams are tested for HGH.

- c) To determine customer opinion of its boarding policy. Southwest Airlines randomly selects 60 flights during a certain week and surveys all passengers on the flights.
- d) A member of Congress wishes to determine her constituency's opinion regarding estate taxes. She divides her constituency into three income classes: low-income households, middle-income households, and upper-income households. She then takes a simple random sample of households from each income class.
- e) In an effort to identify whether an advertising campaign has been effective, a marketing firm conducts a nationwide poll by randomly selecting individuals from a list of known users of the product.
- f) A radio station asks its listeners to call in their opinion regarding the use of U.S. forces in peacekeeping missions.
- g) A farmer divides his orchard into 50 subsections, randomly selects 4, and samples all the trees within the 4 subsections to approximate the yield of this orchard.
- h) A college official divides the student population into five classes: freshman, sophomore, junior, and graduate student. The official takes a simple random sample from each class and asks the members' opinions regarding student services.
- i) Toyota wants to administer a satisfaction survey to its current customers. Using their customer database, the company randomly selects 80 customers and asks them about their level of satisfaction with the company.
- j) To determine her power usage, Dan divides up his day into three parts: morning, afternoon, and evening. He then measures his power usage at 3 randomly selected times during each part of the day.
- k) A newspaper asks its readers to call in their opinion regarding the number of books they have read this month.
- l) Toshiba wants to administer a satisfaction survey to its current customers. Using their customer database, the company randomly selects 80 customers and asks them about their level of satisfaction with the company.
- m) An education researcher randomly selects 48 middle schools and interviews all the teachers at each school.
- n) A market researcher selects 500 drivers under 30 years of age and 500 drivers over 30 years of age.
- o) To avoid working late, a quality control analyst simply inspects the first 100 items produced in a day.

**20.** Determine whether you would take a census or use a sampling to collect data for the study described:

- a) The average credit card debt of the 65 employees of a company
- b) The most popular grocery store among the 40,000 employees of a company

**21.** Determine if the survey question is biased. If the question is biased, suggest a better wording

- a) Why drinking fruit juice good for you?
- b) Why is eating ice cream bad for you?

**22.** A company has been rating television programs for more than 60 years. It uses several sampling procedures, but its main one is to track the viewing patterns of 20,000 households. These contain more than 45,000 people and are chosen to form a cross section of the overall population. The



households represent various locations, ethnic groups, and income brackets. The data gathered from the sample of 20,000 households are used to draw inferences about the population of all households in the U.S.

- a) What strata are used in the sample?
- b) Why is it important to have a stratified sample for these ratings?
- c) Observation studies are sometimes referred to as natural experiments. Explain what this means

23. Some polling agencies ask people to call a telephone number and give their response to a question

- a) What is an advantage of this type of survey?
- b) What is disadvantage of this type of survey?
- c) Identify the sampling technique used.

24. A computer company employs 100 software engineers and hardware engineers. The personnel manager randomly selects 20 of the software engineers and 20 of the hardware and questions them about career opportunities within the company. Does this sampling plan result in a random sample? Simple random sample? Explain.

25. Suppose you are the president of the student government. You wish to conduct a survey to determine that student body's opinion regarding student services. The administration provides you with a list of the names and phone numbers of the 19,935 registered students.

- a) Discuss the procedure you would follow to obtain a simple random sample of 25 students.
- b) Obtain this sample

26. True or False

- a) When taking a systematic random sample of size  $n$ , every group of size  $n$  from the population has the same chance of being selected
- b) A simple random sample is always preferred because it obtains the same information as other sampling plans but requires a smaller sample size.
- c) When conducting a cluster sample, it is better to have fewer cluster with more individuals when the clusters are heterogeneous.
- d) Inferences based on voluntary response samples are generally not reliable.
- e) When obtaining a stratified sample, the number of individuals included within each stratum must be equal.

27. The human resource department at a certain company wants to conduct a survey regarding worker morale. The department has an alphabetical list of all 4502 employees at the company and wants to conduct a systematic sample.

- a) Determine  $k$  if the sample size is 50
- b) Determine the individuals who will be administered the survey. More than one answer is possible.

28. To predict the outcome of a county election, a newspaper obtains a list of all 945,035 registered voters in the county and wants to conduct a systematic sample.

- a) Determine  $k$  if the sample size is 130
- b) Determine the individuals who will be administered the survey. More than one answer is possible.

## Section 1.4 – Design of Experiments

### Describe the Characteristics of an Experiment

#### *Definition*

An **experiment** is a controlled study conducted to determine the effect of varying one or more explanatory variables or **factors** has on a response variable. Any combination of the values of the factors is called a **treatment**.

The **experimental unit** (or **subject**) is a person, object or some other well-defined item upon which a treatment is applied.

A **control group** serves as a baseline treatment that can be used to compare to other treatments.

A **placebo** is an innocuous medication, such as a sugar tablet, that looks, tastes, and smells like the experimental medication.

#### *Definitions*

**Blinding** refers to nondisclosure of the treatment an experimental unit is receiving.

A **single-blind** experiment is one in which the experimental unit (or subject) does not know which treatment he or she is receiving.

A **double-blind** experiment is one in which neither the experimental unit nor the researcher in contact with the experimental unit knows which treatment the experimental unit is receiving.

#### *Example*

Lipitor is a cholesterol-lowering drug made by Pfizer. In the Collaborative Atorvastatin Diabetes Study (CARDS), the effect of Lipitor on cardiovascular disease was assessed in 2838 subjects, ages 40 to 75, with type 2 diabetes, without prior history of cardiovascular disease. In this placebo-controlled, double-blind experiment, subjects were randomly allocated to either Lipitor 10 mg daily (1428) or placebo (1410) and were followed for 4 years. The response variable was the occurrence of any major cardiovascular event.

Lipitor significantly reduced the rate of major cardiovascular events (83 events in the Lipitor group versus 127 events in the placebo group). There were 61 deaths in the Lipitor group versus 82 deaths in the placebo group.

- a) What does it mean for the experiment to be placebo-controlled?
- b) What does it mean for the experiment to be double-blind?
- c) What is the population for which this study applies? What is the sample?
- d) What are the treatments?
- e) What is the response variable? Is it qualitative or quantitative?

#### *Solution*

- a) The placebo is a medication that looks, smells, and tastes like Lipitor. The placebo control group serves as a baseline against which to compare the results from the group receiving Lipitor. The placebo is also used because people tend to behave differently when they are in a study. By having a placebo control group, the effect of this is neutralized.
- b) Since the experiment is double-blind, the subjects, as well as the individual monitoring the subjects, do not know whether the subjects are receiving Lipitor or the placebo. The experiment is double-blind so that the subjects receiving the medication do not behave differently from those receiving the placebo and so the individual monitoring the subjects does not treat those in the Lipitor group differently from those in the placebo group.
- c) The population is individuals from 40 to 75 years of age with type 2 diabetes without a prior history of cardiovascular disease. The sample is the 2838 subjects in the study.
- d) The treatments are 10 mg of Lipitor or a placebo daily.
- e) The response variable is whether the subject had any major cardiovascular event, such as a stroke, or not. It is a qualitative variable.

### ***Example***

The English Department of a community college is considering adopting an online version of the freshman English course. To compare the new online course to the traditional course, an English Department faculty member randomly splits a section of her course. Half of the students receive the traditional course and the other half is given an online version. At the end of the semester, both groups will be given a test to determine which performed better.

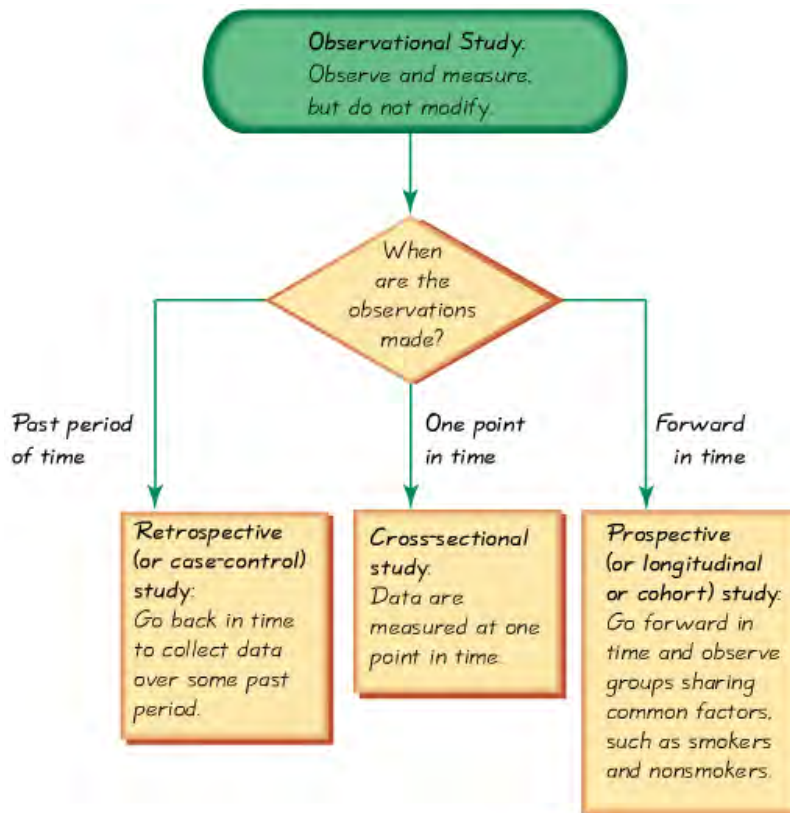
- a) Who are the experimental units?
- b) What is the population for which this study applies?
- c) What are the treatments?
- d) What is the response variable?
- e) Why can't this experiment be conducted with blinding?

### **Solution**

- a) The students in the class
- b) All students who enroll in the class
- c) Traditional vs. online instruction
- d) Exam score
- e) Both the students and instructor know which treatment they are receiving

## Explain the Steps in Designing an Experiment

To *design* an experiment means to describe the overall plan in conducting the experiment.



## Steps in Conducting an Experiment

**Step 1:** Identify the problem to be solved.

- Should be explicit
- Should provide the experimenter direction
- Should identify the response variable and the population to be studied.
- Often referred to as the *claim*.

**Step 2:** Determine the factors that affect the response variable.

- Once the factors are identified, it must be determined which factors are to be fixed at some predetermined level (the control), which factors will be manipulated and which factors will be uncontrolled.

**Step 3:** Determine the number of experimental units.

**Step 4:** Determine the level of the predictor variables

1. **Control:** There are two ways to control the factors.

- a) Fix their level at one predetermined value throughout the experiment. These are variables whose effect on the response variable is not of interest.
- b) Set them at predetermined levels. These are the factors whose effect on the response variable interests us. The combinations of the levels of these factors represent the treatments in the experiment.

2. **Randomize:** Randomize the experimental units to various treatment groups so that the effects of variables whose level cannot be controlled is minimized. The idea is that randomization “averages out” the effect of uncontrolled predictor variables.

**Step 5:** Conduct the Experiment

- a) **Replication** occurs when each treatment is applied to more than one experimental unit. This helps to assure that the effect of a treatment is not due to some characteristic of a single experimental unit. It is recommended that each treatment group have the same number of experimental units.
- b) Collect and process the data by measuring the value of the response variable for each replication. Any difference in the value of the response variable is a result of differences in the level of the treatment.

**Step 6:** Test the claim

- This is the subject of inferential statistics.
- Inferential statistics is a process in which generalizations about a population are made on the basis of results obtained from a sample. Provide a statement regarding the level of confidence in the generalization. Methods of inferential statistics are presented later in the text.

**Randomization** is used when subjects are assigned to different groups through a process of random selection. The logic is to use chance as a way to create two groups that are similar.

**Replication** is the repetition of an experiment on more than one subject. Samples should be large enough so that the erratic behavior that is characteristic of very small samples will not disguise the true effects of different treatments. It is used effectively when there are enough subjects to recognize the differences from different treatments.

Use a sample size that is large enough to let us see the true nature of any effects, and obtain the sample using an appropriate method, such as one based on *randomness*.

**Blinding** is a technique in which the subject doesn't know whether he or she is receiving a treatment or a placebo. Blinding allows us to determine whether the treatment effect is significantly different from a placebo effect, which occurs when an untreated subject reports improvement in symptoms.

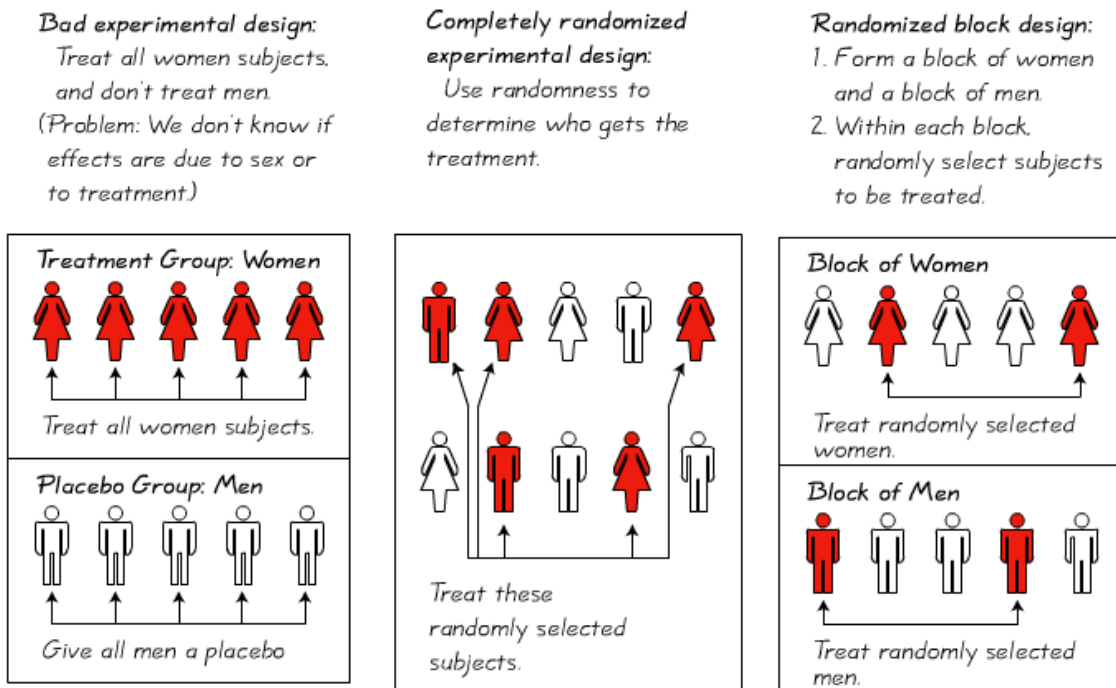
**Double-Blind** : Blinding occurs at two levels:

1. The subject doesn't know whether he or she is receiving the treatment or a placebo
2. The experimenter does not know whether he or she is administering the treatment or placebo

**Completely Randomized Experimental Design** assign subjects to different treatment groups through a process of random selection

**Randomized Block Design** a block is a group of subjects that are similar, but blocks differ in ways that might affect the outcome of the experiment.

**Rigorously Controlled Design** carefully assign subjects to different treatment groups, so that those given each treatment are similar in ways that are important to the experiment



**Matched Pairs Design** compare exactly two treatment groups using subjects matched in pairs that are somehow related or have similar characteristics

### Example

In 1954, a large-scale experiment was designed to test the effectiveness of the Salk vaccine in preventing polio, which had killed or paralyzed thousands of children. In that experiment, 200,745 children were given a treatment consisting of Salk vaccine injections, while a second group of 201,229 children were injected with a placebo that contained no drug. The children being injected did not know whether they were getting the Salk vaccine or the placebo. Children were assigned to the treatment or placebo group a process of random selection, equivalent to flipping a coin. Among the children given the Salk vaccine, 33 later developed paralytic polio, but among the children given a placebo, 115 later develop paralytic polio.

## Explain the Completely Randomized Design

### Definition

A *completely randomized design* is one in which each experimental unit is randomly assigned to a treatment.

### Example

A farmer wishes to determine the optimal level of a new fertilizer on his soybean crop. Design an experiment that will assist him.

### Solution

**Step 1:** The farmer wants to identify the optimal level of fertilizer for growing soybeans. We define optimal as the level that maximizes yield. So the response variable will be crop yield.

**Step 2:** Some factors that affect crop yield are fertilizer, precipitation, sunlight, method of tilling the soil, type of soil, plant, and temperature.

**Step 3:** In this experiment, we will plant 60 soybean plants (experimental units)

**Step 4:** We list the factors and their levels.

**Fertilizer:** This factor will be controlled and set at three levels. We wish to measure the effect of varying the level of this variable on the response variable, yield. We will set the treatments (level of fertilizer) as follows.

**Treatment A:** 20 soybean plants receive no fertilizer.

**Treatment B:** 20 soybean plants receive 2 teaspoons of fertilizer per gallon of water every 2 weeks.

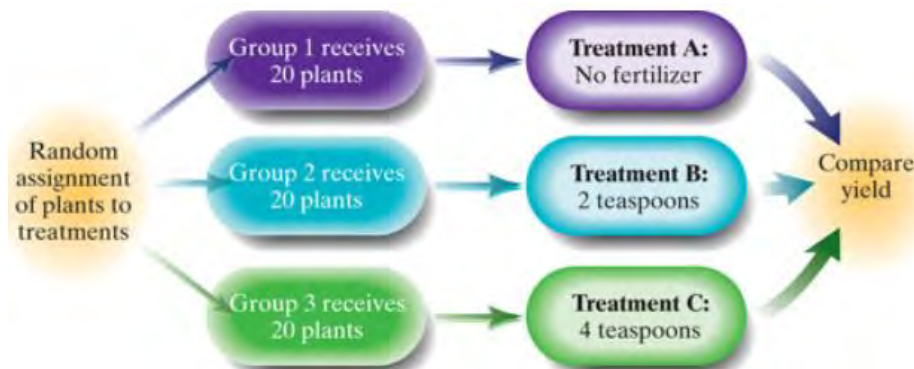
**Treatment C:** 20 soybean plants receive 4 teaspoons of fertilizer per gallon of water every 2 weeks.



**Step 5: a)** We need to assign each plant to a treatment group. First, we will number the plants from 1 to 60 and randomly generate 20 numbers. The plants corresponding to these numbers get treatment *A*. Next we number the remaining plants 1 to 40 and randomly generate 20 numbers. The plants corresponding to these numbers get the treatment *B*. The remaining plants get treatment *C*. Now we till the soil, plant the soybean plants, and fertilizer according to the schedule prescribed.

**b)** At the end of the growing season, we determine the crop yield for each plant.

**Step 6:** We determine any difference in yield among the three treatment groups.





## Example

The octane of fuel is a measure of its resistance to detonation with a higher number indicating higher resistance. An engineer wants to know whether the level of octane in gasoline affects the gas mileage of an automobile. Assist the engineer in designing an experiment.

### Solution

**Step 1:** The response variable is miles per gallon.

**Step 2:** Factors that affect miles per gallon:

Engine size, outside temperature, driving style, driving conditions, characteristics of car

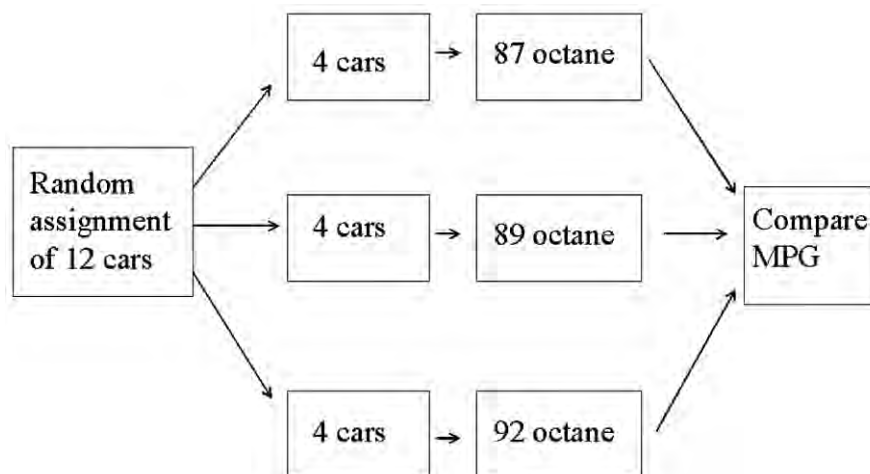
**Step 3:** We will use 12 cars all of the same model and year.

**Step 4:** We list the variables and their level

- Octane level - manipulated at 3 levels. Treatment A: 87 octane, Treatment B: 89 octane, Treatment C: 92 octane
- Engine size - fixed
- Temperature - uncontrolled, but will be the same for all 12 cars.
- Driving style/conditions - all 12 cars will be driven under the same conditions on a closed track - fixed.
- Other characteristics of car - all 12 cars will be the same model year, however, there is probably variation from car to car. To account for this, we randomly assign the cars to the octane level.

**Step 5:** Randomly assign 4 cars to the 87 octane, 4 cars to the 89 octane, and 4 cars to the 92 octane. Give each car 3 gallons of gasoline. Drive the cars until they run out of gas. Compute the miles per gallon.

**Step 6:** Determine whether any differences exist in miles per gallon.



## Explain the Matched-Pairs Design

### *Definition*

A **matched-pairs design** is an experimental design in which the experimental units are paired up. The pairs are matched up so that they are somehow related (that is, the same person before and after a treatment, twins, husband and wife, same geographical location, and so on). There are only two levels of treatment in a matched-pairs design.

### *Example*

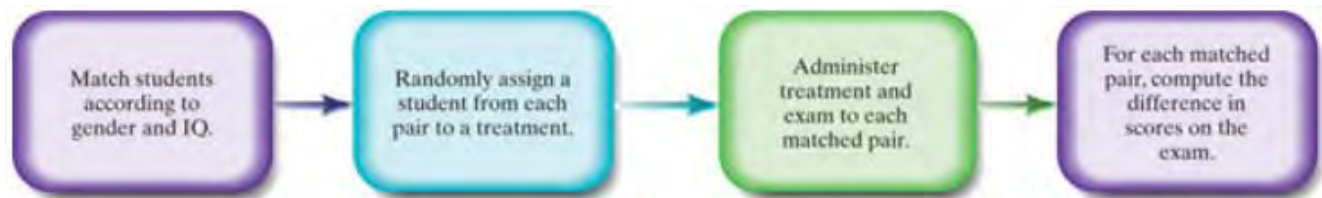
An educational psychologist want to determine listening to music has an effect on a student's ability to learn. Design an experiment to help the psychologist answer the question.

### *Solution*

We match students according to IQ and gender. For example, we match two females with IQs in the 110 to 115 range.

For each pair of students, we flip a coin to determine which student is assigned the treatment of a quiet room or a room with music playing in the background.

Each student will be given a statistics textbook. After 2 hours the students will enter a testing center and take a short quiz on the material in the section. We compute the difference in the scores of each matched pair. Any differences in scores will be attributed to the treatment.



## Explain the Randomized Block Design

Grouping similar (homogeneous) experimental units together and then randomizing the experimental units within each group to a treatment is called blocking. Each group of homogeneous individuals is called a **block**.

**Confounding** occurs when the effect of two factors (explanatory variables) on the response variable cannot be distinguished.

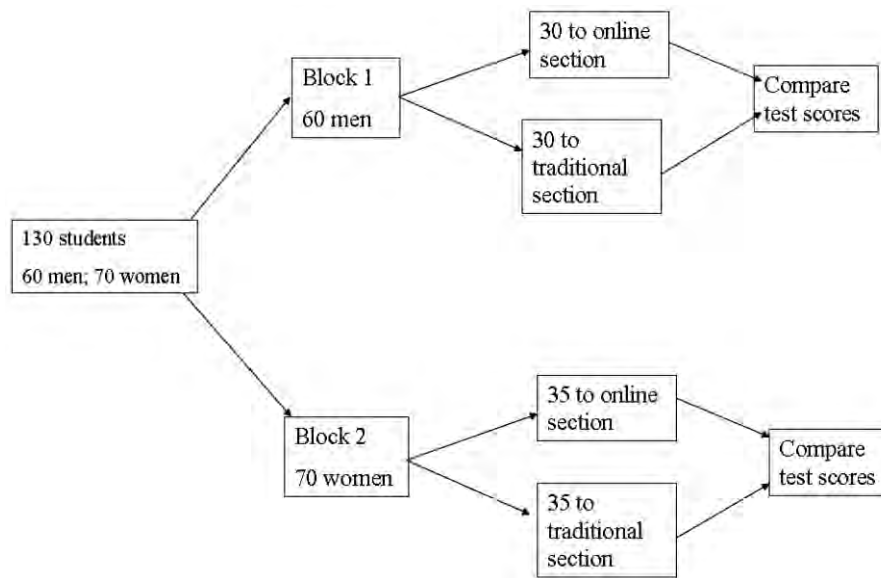
A **randomized block design** is used when the experimental units are divided into homogeneous groups called blocks. Within each block, the experimental units are randomly assigned to treatments.

### *Example*

The English Department is considering adopting an online version of the freshman English course. After some deliberation, the English Department thinks that there may be a difference in the performance of the men and women in the traditional and online courses. To accommodate any potential differences, they randomly assign half the 60 men to each of the two courses and they do the same for the 70 women.

### Solution

This is a randomized block design where gender forms the block. This way, gender will not play a role in the value of the response variable, test score. We do not compare test results across gender.



### Summary

Three very important considerations in the design of experiments are the following:

1. Use *randomization* to assign subjects to different groups
2. Use replication by repeating the experiment on enough subjects so that effects of treatment or other factors can be clearly seen.
3. *Control the effects of variables* by using such techniques as blinding and a completely randomized experimental design

### Errors

No matter how well you plan and execute the sample collection process, there is likely to be some error in the results.

**Sampling error** the difference between a sample result and the true population result; such an error results from chance sample fluctuations

**Nonsampling error** sample data incorrectly collected, recorded, or analyzed (such as by selecting a biased sample, using a defective instrument, or copying the data incorrectly)

## ***Exercises***     **Section 1.4 – Design of Experiments**

1. A school psychologist wants to test the effectiveness of a new method for teaching reading. She recruits 500 first-grade students in District 203 and randomly divides them into two groups. Group 1 is taught by means of the new method, while group 2 is taught by traditional methods. The same teacher is assigned to teach both groups. At the end of the year, an achievement test is administered and the results of the two groups are compared.
  - a) What is the response variable in this experiment?
  - b) Think of some of the factors in the study. How are they controlled?
  - c) What are the treatments? How many treatments are there?
  - d) How are the factors that are not controlled dealt with?
  - e) Which group serves as the control group?
  - f) What type of experimental design is this?
  - g) Identify the subjects.
  
2. A pharmaceutical company has developed an experimental drug meant to relieve symptoms associated with the common cold. The company identifies 300 adult males 25 to 29 years old who have a common cold and randomly divides them into 2 groups. Group 1 is given the experimental drug, while group 2 is given a placebo. After 1 week of treatment, the proportions of each group that still have cold symptoms are compared.
  - a) What is the response variable in this experiment?
  - b) Think of some of the factors in the study. How are they controlled?
  - c) What are the treatments? How many treatments are there?
  - d) How are the factors that are not controlled dealt with?
  - e) What type of experimental design is this?
  - f) Identify the subjects.
  
3. Researchers wanted to compare the effectiveness and safety of an extract of St. John's wort with placebo in outpatients with major depression. To do this, they recruited 200 adult outpatients diagnosed as having major depression and having a baseline Hamilton Rating Scale for Depression (HAM-D) score of at least 20. Participants were randomly assigned to receive either St. John's wort extract, 900 *mg* per day (*mg/d*) for a weeks, increased to 1200 *mg/d* in the absence of an adequate response thereafter, or a placebo for 8 weeks. The response variable was the change on the HAM-D over the treatment period. After analysis of the data, it was concluded that St. John's wort was not effective for treatment of major depression.
  - a) What type of experimental design is this?
  - b) What is the population that is being studied?
  - c) What is the response variable in this study?
  - d) What are the treatments?
  - e) Identify the experimental units.
  - f) What is the control group in this study?

4. Researchers wanted to evaluate whether ginkgo, an over-the-counter herb marketed as enhancing memory, improves memory in elderly adults as measured by objective tests. To do this, they recruited 96 men and 132 women older than 60 years and in good health. Participants were randomly assigned to receive ginkgo, 40 *mg* 3 times per day, or a matching placebo. The measure of memory improvement was determined by a standardized test of learning and memory. After 6 weeks of treatment, the data indicated that ginkgo did not increase performance on standard tests of learning, memory, attention, and concentration. These data suggest that, when taken following the manufacturer's instructions, ginkgo provides no measurable increase in memory or related cognitive function to adults with healthy cognitive function.
- a)* What type of experimental design is this?
  - b)* What is the population being studied?
  - c)* What is the response variable in this study?
  - d)* What is the factor that is set to predetermined levels? What are the treatments?
  - e)* Identify the experimental units.
  - f)* What is the control group in this study?

## Section 1.5 – Organizing Qualitative Data

When data is collected from a survey or designed experiment, they must be organized into a manageable form. Data that is not organized is referred to as *raw data*.

### Ways to Organize Data

- Tables
- Graphs
- Numerical Summaries

#### Definition

A *frequency distribution* (or *frequency table*) shows how a data set is partitioned among all of several categories (or classes) by listing all of the categories along with the number of data values in each of the categories.

#### Example

Consider pulse rate measurements (in beats per minute) obtained from a simple random sample of 40 males and another simple random sample of 40 females, with the results listed in the table below.

**Pulse Rates (*beats per minute*) of Females and Males**

<i>Females</i>																			
76	72	88	60	72	68	80	64	68	68	80	76	68	72	93	72	68	72	64	80
64	80	76	76	76	80	104	88	60	76	72	72	88	80	60	72	88	88	124	64
<i>Males</i>																			
68	64	88	72	64	72	60	88	76	60	96	72	56	64	60	64	84	76	84	88
72	56	68	64	60	68	60	60	56	84	72	84	88	56	64	56	56	60	64	72

The frequency distribution summarizing the pulse rate of females listed in table below.

**Pulse Rates of Females**

<i>Pulse Rate</i>	<i>Frequency</i>
60 – 69	12
70 – 79	14
80 – 89	11
90 – 99	1
100 – 109	1
110 – 119	0
120 – 129	1

The frequency for a particular class is the number of original values that fall into that class. That is the frequency of 12, indicating that 12 of the original pulse rates are between 60 and 69 beats per minute.

## Relative Frequency Distribution

The *relative frequency* is the proportion (or percent) of observations within a category and is found using the formula:

$$\text{relative frequency} = \frac{\text{class frequency}}{\text{sum of all frequency}} \quad \text{percentage frequency} = \frac{\text{class frequency}}{\text{sum of all frequency}} \times 100\%$$

A variation of the basic frequency distribution is a *relative frequency distribution*.

<i>Pulse Rate</i>	<i>Frequency</i>	<i>Relative Frequency</i>	<i>Relative Frequency %</i>
60 – 69	12	$\frac{12}{40} = 0.3$	$\frac{12}{40} \times 100\% = 30\%$
70 – 79	14	$\frac{14}{40} = 0.35$	$\frac{14}{40} \times 100\% = 35\%$
80 – 89	11	$\frac{11}{40} = 0.275$	$\frac{11}{40} \times 100\% = 27.5\%$
90 – 99	1	$\frac{1}{40} = 0.025$	$\frac{1}{40} \times 100\% = 2.5\%$
100 – 109	1	$\frac{1}{40} = 0.025$	$\frac{1}{40} \times 100\% = 2.5\%$
110 – 119	0	0	0
120 – 129	1	$\frac{1}{40} = 0.025$	$\frac{1}{40} \times 100\% = 2.5\%$
	<b>40</b>		

### Example

The data below represent the color of M&Ms in a bag of plain M&Ms.

brown, brown, yellow, red, red, red, brown, orange, blue, green, blue, brown, yellow, yellow, brown, red, red, brown, brown, brown, green, blue, green, orange, orange, yellow, yellow, yellow, red, brown, red, brown, orange, green, red, brown, yellow, orange, red, green, yellow, yellow, brown, yellow, orange

Construct a frequency distribution and a relative frequency distribution of the color of plain M&Ms.

### Solution

<i>Color</i>	<i>Tally</i>	<i>Frequency</i>	<i>Relative Frequency</i>
Brown		12	$\frac{12}{45} \approx 0.2667$
Yellow		10	$\frac{10}{45} \approx 0.2222$
Red		9	$\frac{9}{45} = 0.2$
Orange		6	$\frac{6}{45} \approx 0.1333$
Blue		3	$\frac{3}{45} \approx 0.0667$
Green		5	$\frac{5}{45} \approx 0.1111$
		<b>45</b>	

## Construct Bar Graphs

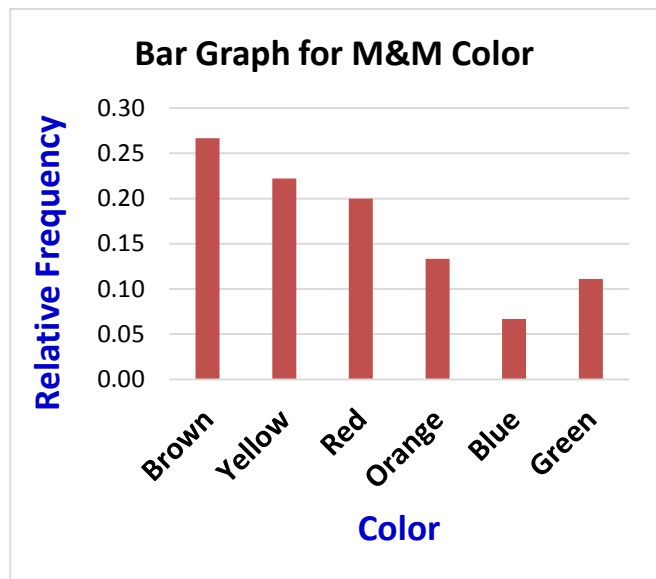
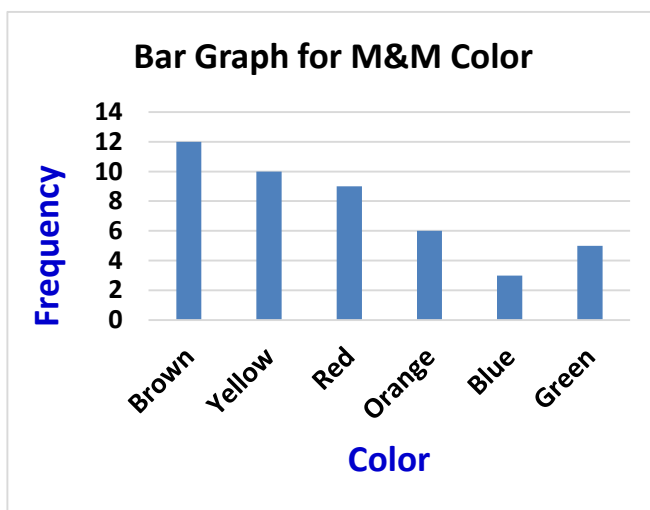
A **bar graph** is constructed by labeling each category of data on either the horizontal or vertical axis and the frequency or relative frequency of the category on the other axis. Rectangles of equal width are drawn for each category. The height of each rectangle represents the category's frequency or relative frequency.

### Example

Use the M&M data to construct

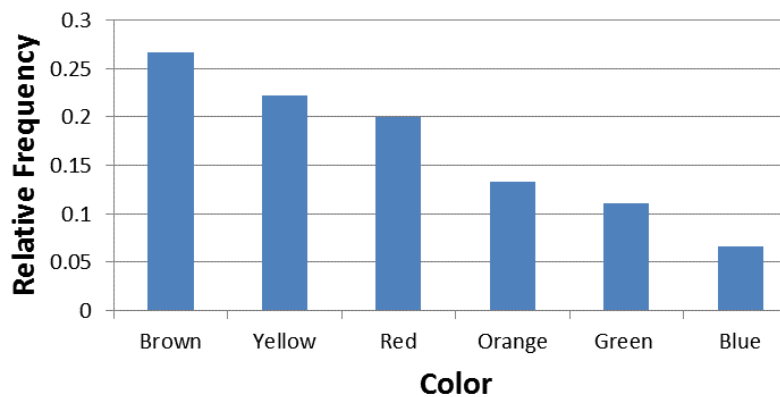
- a frequency bar graph and
- a relative frequency bar graph.

### Solution



A **Pareto chart** is a bar graph where the bars are drawn in decreasing order of frequency or relative frequency.

### *Pareto Chart* Colors of M&Ms





## Construct Pie Charts

A *pie chart* is a circle divided into sectors. Each sector represents a category of data. The area of each sector is proportional to the frequency of the category.

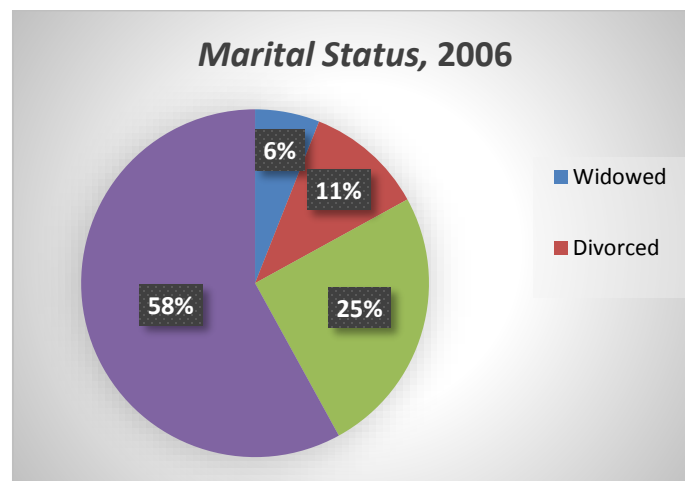
### Example

The following data represent the marital status (in millions) of U.S. residents 18 years of age or older in 2006. Draw a pie chart of the data.

<i>Marital Status</i>	<i>Frequency</i>
Never married	55.3
Married	127.7
Widowed	13.9
Divorced	22.8

### Solution

<i>Marital Status</i>	<i>Frequency</i>	
Never married	55.3	$\frac{55.3}{219.7} \times 100 \approx 25\%$
Married	127.7	$\frac{127.7}{219.7} \times 100 \approx 58\%$
Widowed	13.9	$\frac{13.9}{219.7} \times 100 \approx 6\%$
Divorced	22.8	$\frac{22.8}{219.7} \times 100 \approx 11\%$
	<b>219.7</b>	



## Normal Distribution

- The *frequencies* start low, then increase to one or two high frequencies, then decrease to a low frequency.
- The distribution is approximately symmetric, with frequencies preceding the maximum being roughly a mirror image of those that follow the maximum.

### Example

IQ scores from 1000 adults were randomly selected. The results are summarized in the frequency distribution table

<i><b>IQ Score</b></i>	<i><b>Frequency</b></i>	<i><b>Normal Distribution</b></i>
50 – 69	24	← Frequencies start low
70 – 89	228	
90 – 109	490	← Increase to a maximum, ...
110 – 129	232	
130 – 149	26	← Decrease to become low again

The frequencies start low, then increase to a maximum frequency of 490, then decrease to low frequencies. Also, the frequencies are roughly symmetric about the maximum frequency of 490. It appears that the distribution is approximately a normal distribution.



19	123.8	124.5	123.7		39	123.8	124.3	122.9
20	123.8	124.6	123.7		40	123.8	124.0	123.0

5. As part of the Garbage Project at the University of Arizona, the discarded garbage for 62 households was analyzed. Refers to the 62 weights from table below and construct a frequency distribution. Begin with a lower class of 1.00 lb., and use a class width of 4.00 lb. Do the weights of discarded paper appear to have a normal distribution?

2.41	11.08	9.45	5.88
7.57	12.43	12.32	8.26
9.55	6.05	20.12	12.45
8.82	13.61	7.72	10.58
8.72	6.98	6.16	5.87
6.96	14.33	7.98	8.78
6.83	13.31	9.64	11.03
11.42	3.27	8.08	12.29
16.08	6.67	10.99	20.58
6.38	17.65	13.11	12.56
13.05	12.73	3.26	9.92
11.36	9.83	1.65	3.45
15.09	16.39	10	9.09
2.8	6.33	8.96	3.69
6.44	9.19	9.46	2.61
5.86	9.41		

6. a) Refer to the data below for the FICO credit rating scores. Construct a frequency distribution beginning with a lower class limit of 400, and use a class width of 50. Does the result appear to have a normal distribution? Why or why not?

708	713	781	809	797	793	711	681	768	611	698	729	829
836	768	532	657	559	741	792	701	753	745	681	594	744
598	693	743	444	502	739	755	835	714	517	787	706	752
714	497	636	637	797	568	714	618	830	579	818	722	783
751	731	850	591	802	756	689	789	654	617	849	604	630
628	692	779	756	782	760	503	784	798	611	709	661	579
591	834	694	795	660	651	696	638	697	732	796	753	782
635	795	519	682	824	603	709	777	664				

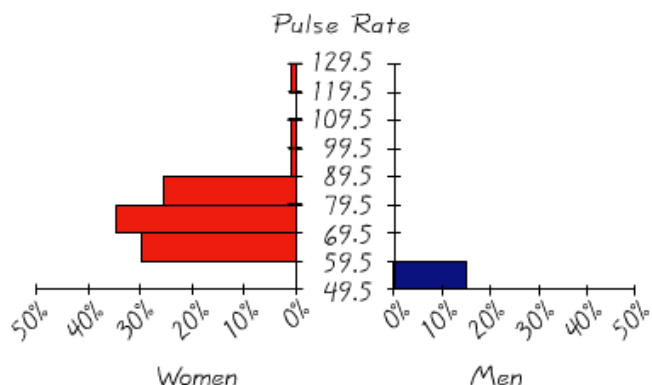
- b) Use the table to construct a histogram. Does the result appear to be normal distribution? Why or why not?

7. a) Refer to the data in the table below. Construct a frequency distribution. Begin with lower class limit of 6.0000 g, and use a class width of 0.0500 g.

6.2771	6.2371	6.1501	6.0002	6.1275	6.2151	6.1947	6.1940
6.2866	6.0760	6.1426	6.3415	6.1309	6.2412	6.2130	6.0257
6.1442	6.1073	6.1181	6.1352	6.2821	6.2647	6.1787	6.1719
6.2908	6.1661	6.2674	6.2718	6.1949	6.2465	6.1095	6.3278
6.3172	6.1487	6.0829	6.1423	6.1970	6.2441	6.0775	6.3669

- b) Use the table to construct a histogram.

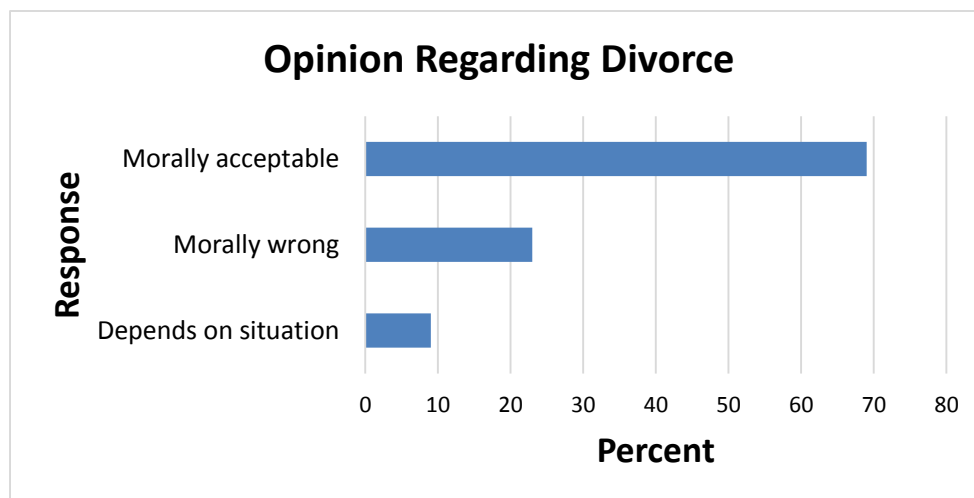
8. When using histograms to compare two data sets, it is sometimes difficult to make comparisons by looking back and forth between the two histograms. A back-to-back relative frequencies histogram uses a format that makes the comparison much easier. Instead of frequencies, we should use relative frequencies (percentages or proportions) so that the comparisons are not distorted by different sample sizes. Complete the back-to-back relative frequency histograms shown below by using the data below. Then use the result to compare the two data sets.



**Pulse Rates (*beats per minute*) of Females and Males**

<i>Females</i>																			
76	72	88	60	72	68	80	64	68	68	80	76	68	72	93	72	68	72	64	80
64	80	76	76	76	80	104	88	60	76	72	72	88	80	60	72	88	88	124	64
<i>Males</i>																			
68	64	88	72	64	72	60	88	76	60	96	72	56	64	60	64	84	76	84	88
72	56	68	64	60	68	60	60	56	84	72	84	88	56	64	56	56	60	64	72

9. The following graph represents the results of a survey, by Gallup in May 2010, in which a random sample of adult Americans was asked, “Please tell me whether you personally believe that in general divorce is morally accepted or morally wrong.”



- What percent of the respondents believe divorce is morally acceptable?
  - If there were 240 million adult Americans, how many believe that divorce is morally wrong?
  - If Gallup claimed that the results of the survey indicate that 8% of adult Americans believe that divorce is acceptable in certain situations, would you say this statement is descriptive or inferential? Why?
10. In a national survey conducted by the Centers for Disease Control to determine health-risk behaviors among college students, college students were asked, “How often do you wear a seat belt when driving a car?” The frequencies were as follow:

<i>Response</i>	<i>Frequency</i>
I do not drive a car	249
Never	118
Rarely	249
Sometimes	345
Most of the time	716
Always	3093

- Construct a relative frequency distribution
- What percentage of respondents answered “Always”?
- What percentage of respondents answered “Never” or “Rarely”?
- Construct a frequency bar graph.
- Construct a relative frequency bar graph.
- Construct a pie chart
- Suppose that a representative from the Centers for Disease Control says, “2.5% of the college students in this survey responded that they never wear a seat belt.” Is this a descriptive or inferential statement?

11. A phlebotomist draws the blood of a random sample of 50 patients and determines their blood types as shown.

<i>O</i>	<i>B</i>	<i>AB</i>	<i>O</i>	<i>AB</i>	<i>O</i>	<i>O</i>	<i>O</i>	<i>O</i>	<i>O</i>
<i>O</i>	<i>O</i>	<i>B</i>	<i>O</i>	<i>O</i>	<i>A</i>	<i>A</i>	<i>B</i>	<i>O</i>	<i>A</i>
<i>A</i>	<i>B</i>	<i>A</i>	<i>A</i>	<i>A</i>	<i>A</i>	<i>O</i>	<i>A</i>	<i>O</i>	<i>O</i>
<i>A</i>	<i>A</i>	<i>B</i>	<i>A</i>	<i>B</i>	<i>O</i>	<i>AB</i>	<i>A</i>	<i>A</i>	<i>A</i>
<i>O</i>	<i>O</i>	<i>AB</i>	<i>O</i>	<i>A</i>	<i>A</i>	<i>A</i>	<i>O</i>	<i>O</i>	<i>O</i>

- Construct a frequency distribution
- Construct a relative frequency distribution
- According to the data, which blood type is most common?
- According to the data, which blood type is least common?
- Use the results of the sample to conjecture the percentage of the population that has type *O* blood. Is this an example of descriptive or inferential statistics?
- Contact a local hospital and ask them the percentage of the population that us blood type *O*. Why might the results differ?
- Draw a frequency bar graph
- Draw a relative frequency bar graph
- Draw a pie chart

## Section 1.6 – Additional Displays

### Discrete Data

The first step in summarizing quantitative data is to determine whether the data are discrete or continuous. If the data are discrete and there are relatively few different values of the variable, the categories of data (classes) will be the observations (as in qualitative data). If the data are discrete, but there are many different values of the variables, or if the data are continuous, the categories of data (the *classes*) must be created using intervals of numbers.

### Construct Histograms of Discrete Data

#### Definitions

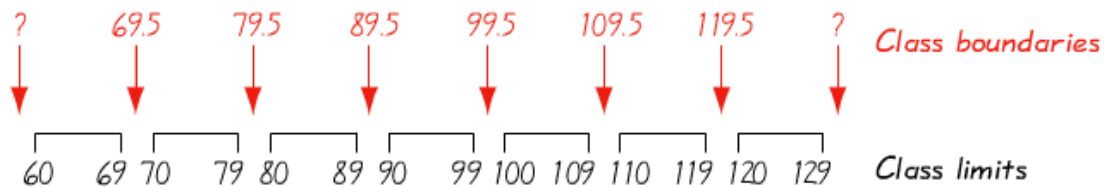
A **histogram** is constructed by drawing rectangles for each class of data. The height of each rectangle is the frequency or relative frequency of the class. The width of each rectangle is the same and the rectangles touch each other.

**Classes** are categories into which data are grouped. When a data set consists of a large number of different discrete data values or when a data set consists of continuous data, we must create classes by using intervals of numbers.

**Lower class limits** are the smallest numbers that can belong to the different classes (60, 70, 80, 90, 100, 110, and 120) (from previous table).

**Upper class limits** are the largest numbers that can belong to the different classes (table has upper class limits of 69, 79, 89, 99, 109, 119, 129.)

**Class boundaries** are the numbers used to separate the classes, but without the gaps created by the class limits. So the complete list of class boundaries is 59.5, 69.5, 79.5, 89.5, 99.5, 109.5, 119.5, 129.5.



**Class midpoints** are the values in the middle of the classes. From the table: 64.5, 74.5, 84.5, 94.5, 104.5, 114.5, and 124.5.) Each class midpoint is found by adding the lower class limits to the upper class limits and dividing the sum by 2.

**Class width** is the difference between two consecutive lower class limits or two consecutive lower class boundaries in a frequency distribution. (Table uses a class width of 10.)

### ***Example***

The following data represent the time between eruptions (in seconds) for a random sample of 45 eruptions at the Old Faithful Geyser in Wyoming. Construct a frequency and relative frequency distribution of the data.

728	678	723	735	703
730	722	708	714	713
726	716	736	719	672
698	702	738	725	711
721	703	735	699	695
722	718	695	702	731
700	703	706	733	726
720	723	711	696	695
729	699	714	700	718

### **Solution**

The smallest data value is 672 and the largest data value is 738. We will create the classes so that the lower class limit of the first class is 670 and the class width is 10 and obtain the following classes:

<i>Time</i>	<i>Frequency</i>	<i>Relative Frequency</i>
670 – 679	2	$\frac{2}{45} = 0.044$
680 – 689	0	0
690 – 699	7	$\frac{7}{45} \approx 0.1556$
700 – 709	9	$\frac{9}{45} = 0.2$
710 – 719	9	$\frac{9}{45} = 0.2$
720 – 729	11	$\frac{11}{45} \approx 0.2444$
730 - 739	7	$\frac{7}{45} \approx 0.1556$

- The choices of the lower class limit of the first class and the class width were rather arbitrary.
- There is not one correct frequency distribution for a particular set of data.
- However, some frequency distributions can better illustrate patterns within the data than others. So constructing frequency distributions is somewhat of an art form.
- Use the distribution that seems to provide the best overall summary of the data.

### **Guidelines for Determining the Lower Class Limit of the First Class and Class Width**

Choosing the Lower Class Limit of the First Class.

Choose the smallest observation in the data set or a convenient number slightly lower than the smallest observation in the data set.



## Procedure for Constructing a Frequency Distribution

The steps for constructing the frequency distributions are as follows:

1. Determine the number of classes. The number of classes should be between 5 and 20, and the number you select might be affected by the convenience of using round numbers.
2. Calculate the class width.

$$\text{Class width} \approx \frac{(\text{maximum data value}) - (\text{minimum data value})}{\text{number of classes}}$$

Round this result to get a convenient number (usually round up). If necessary, change the number of classes so that they use convenient values.

3. Choose either the minimum data value or a convenient value below the minimum data value as the first lower class limit.
4. Using the first lower class limit and the class width, list the other lower class limits. (Add the class width to the first lower class limit to get the second lower class limit. Add the class width to the second lower class limit to get the third lower class limit, and so on.)
5. List the lower class limits in a vertical column and then enter the upper class limits.
6. Take each individual data value and put a tally mark in the appropriate class. Add the tally marks to find the total frequency for each class.

### Example

Using the pulse rate of females in previous table, follow the above procedure to construct the frequency distribution

**Pulse Rates of Females**

<i>Pulse Rate</i>	<i>Frequency</i>
60 – 69	12
70 – 79	14
80 – 89	11
90 – 99	1
100 – 109	1
110 – 119	0
120 – 129	1

### Solution

**Step 1:** The number of desired classes is 7.

**Step 2:**  $\text{Class width} \approx \frac{124 - 60}{7} \approx 9.142857 \approx 10$

**Step 3:** The minimum data value is 60 and it is also a convenient number.

**Step 4:** Add the class width of 10 to 60 to get 70 (second lower class limit)  
 $70 + 10 = 80$ ,  $80 + 10 = 90$ ,  $90 + 10 = 100$ ,  $100 + 10 = 110$ , and  $110 + 10 = 120$ .

**Step 5:** List the lower class limits vertically as shown in the margin. From this list, we identify the corresponding upper class limits as 69, 79, 89, 99, 109, 119, and 129.

**Step 6:** Enter a tally mark for each data value in the appropriate class. Then add the tally marks to find the frequencies.

60 –
70 –
80 –
90 –
100 –
110 –
120 –

## Cumulative Frequency Distribution

The *cumulative frequency* for a class is the sum of the frequencies for that class and all previous classes. The *cumulative frequency distribution* based on the frequency distribution.

<i>Pulse Rate</i>	<i>Frequency</i>	<i>Cumulative Frequency</i>
60 – 69	12	12
70 – 79	14	$12 + 14 = 26$
80 – 89	11	$26 + 11 = 37$
90 – 99	1	$37 + 1 = 38$
100 – 109	1	$38 + 1 = 39$
110 – 119	0	$39 + 0 = 39$
120 – 129	1	$39 + 1 = 40$

## Dot Plots

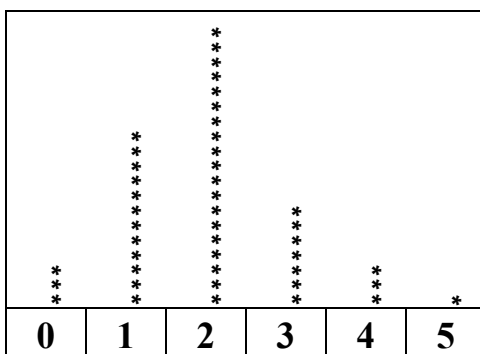
A dotplot consists of a graph in which each data value is plotted as a *point* (or *dot*) along a scale of values. Dots representing equal values are stacked.

### Example

The following data represent the number of available cars in a household based on a random sample of 45 households. Draw a dot plot of the data.

3	0	1	2	1	1	2	0	0
4	2	2	2	1	2	4	2	1
1	1	3	2	4	1	2	2	3
3	3	2	1	2	2	2	3	1
2	3	2	1	2	2	5	2	1

### Solution



## Stemplots

A **stemplot** (or **stem-and-leaf plot**) represents quantitative data by separating each value into two parts: the **stem** (such as the *leftmost digit*) and the **leaf** (such as the *rightmost digit*)

For **example**, a data value of 147 would have 14 as the stem and 7 as the leaf.

**Example** of pulse

<b>Stem</b> (tens)	<b>Leaves</b> (units)	
6	000444488888	← Data values are 60, 60, 60, 64, 64, ..., 68
7	22222222666666	← Data values are 72, 72, ...
8	00000088888	← Data values are 80, 80, ....
9	6	← Data value is 96
10	4	← Data value is 104
11		
12	4	← Data value is 124

## Construction of a Stem-and-leaf Plot

**Step 1** The stem of a data value will consist of the digits to the left of the right- most digit. The leaf of a data value will be the rightmost digit.

**Step 2** Write the stems in a vertical column in increasing order. Draw a vertical line to the right of the stems.

**Step 3** Write each leaf corresponding to the stems to the right of the vertical line.

**Step 4** Within each stem, rearrange the leaves in ascending order, title the plot, and include a legend to indicate what the values represent.

## Advantage of Stem-and-Leaf Diagrams over Histograms

Once a frequency distribution or histogram of continuous data is created, the raw data is lost (unless reported with the frequency distribution), however, the raw data can be retrieved from the stem-and-leaf plot.

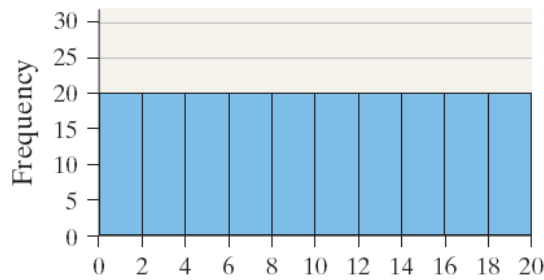
## Identify the Shape of a Distribution

**Uniform distribution** the frequency of each value of the variable is evenly spread out across the values of the variable

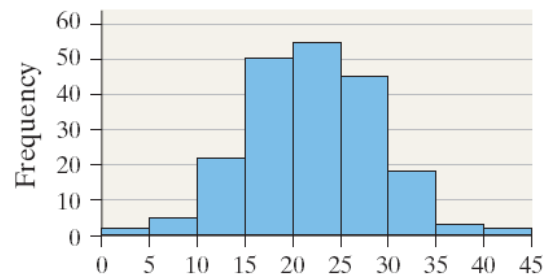
**Bell-shaped distribution** the highest frequency occurs in the middle and frequencies tail off to the left and right of the middle

**Skewed right** the tail to the right of the peak is longer than the tail to the left of the peak

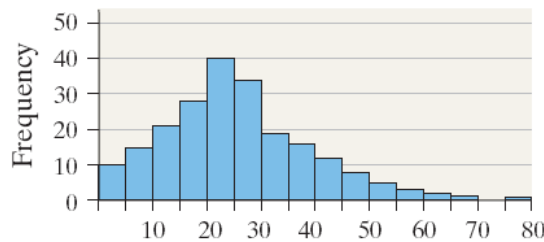
**Skewed left** the tail to the left of the peak is longer than the tail to the right of the peak.



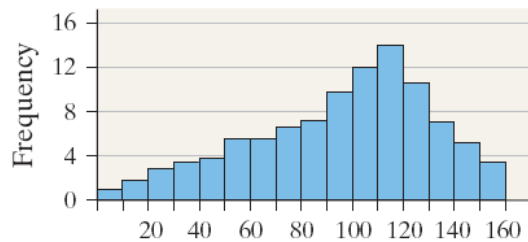
(a) Uniform (symmetric)



(b) Bell-shaped (symmetric)



(c) Skewed Right



(d) Skewed Left

## Frequency Polygon

A **class midpoint** is the sum of consecutive lower class limits divided by 2.

A **frequency polygon** is a graph that uses points, connected by line segments, to represent the frequencies for the classes. It is constructed by plotting a point above each class midpoint on a horizontal axis at a height equal to the frequency of the class. Next, line segments are drawn connecting consecutive points. Two additional line segments are drawn connecting each end of the graph with the horizontal axis.

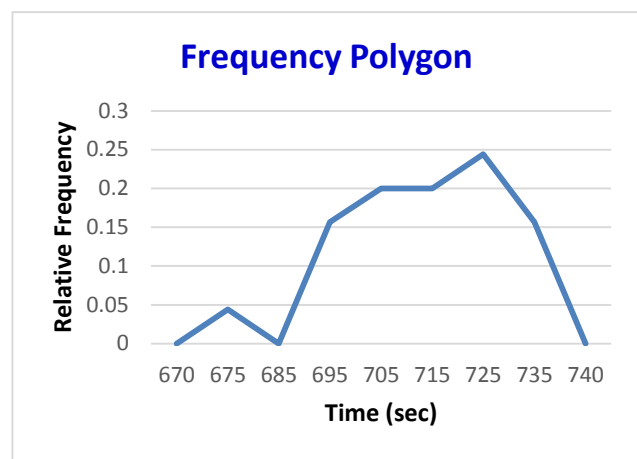
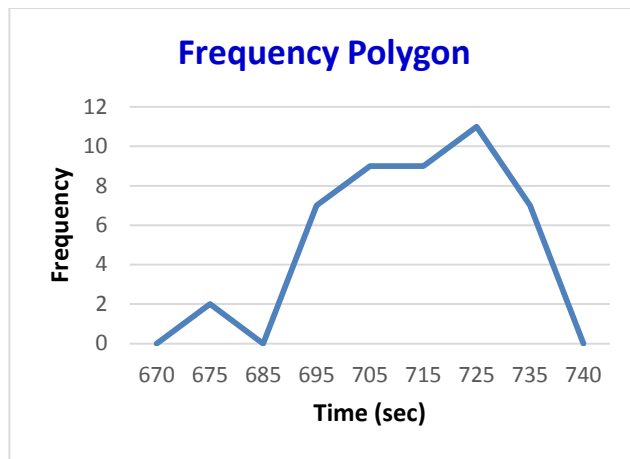
### Example

The following data represent the time between eruptions (in seconds) for a random sample of 45 eruptions at the Old Faithful Geyser in Wyoming. Construct a frequency and relative frequency polygon of the data.

728	711	703	678	695	723	735	700	703
730	723	706	722	718	708	714	714	713
726	720	700	716	702	736	719	699	672
698	726	731	702	733	738	725	729	711
721	696	718	703	722	735	699	695	695

### Solution

<i>Time</i>	<i>Class Midpoint</i>	<i>Frequency</i>	<i>Relative Frequency</i>
670 – 679	675	2	0.044
680 – 689	685	0	0
690 – 699	695	7	0.1556
700 – 709	705	9	0.2
710 – 719	715	9	0.2
720 – 729	725	11	0.2444
730 - 739	735	7	0.1556



## Create Cumulative Frequency and Relative Frequency Tables

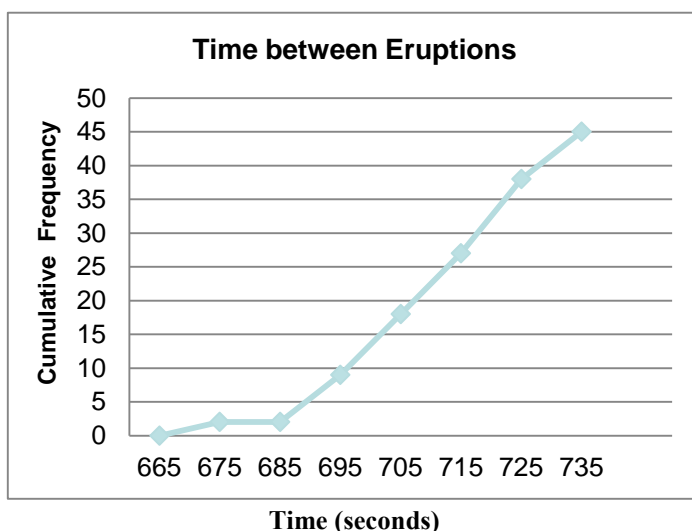
A **cumulative frequency distribution** displays the aggregate frequency of the category. In other words, for discrete data, it displays the total number of observations less than or equal to the category. For continuous data, it displays the total number of observations less than or equal to the upper class limit of a class.

A **cumulative relative frequency distribution** displays the proportion (or percentage) of observations less than or equal to the category for discrete data and the proportion (or percentage) of observations less than or equal to the upper class limit for continuous data.

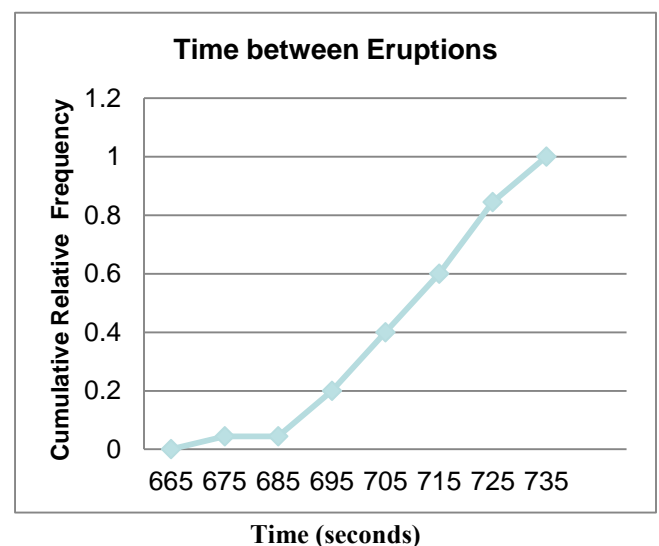
## Ogive (“oh-jive”)

**Ogives** are useful for determining the number of values below some particular value. An **ogive** is a line graph that depicts *cumulative* frequencies. An ogive uses class boundaries along the horizontal scale, and cumulative frequencies along the vertical scale.

**Frequency Ogive**

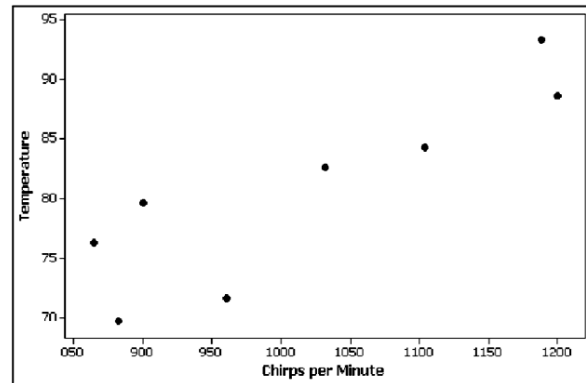


**Relative Frequency Ogive**



## Scatter Plot (or Scatter Diagram)

A plot of paired  $(x,y)$  data with a horizontal  $x$ -axis and a vertical  $y$ -axis. Used to determine whether there is a relationship between the two variables



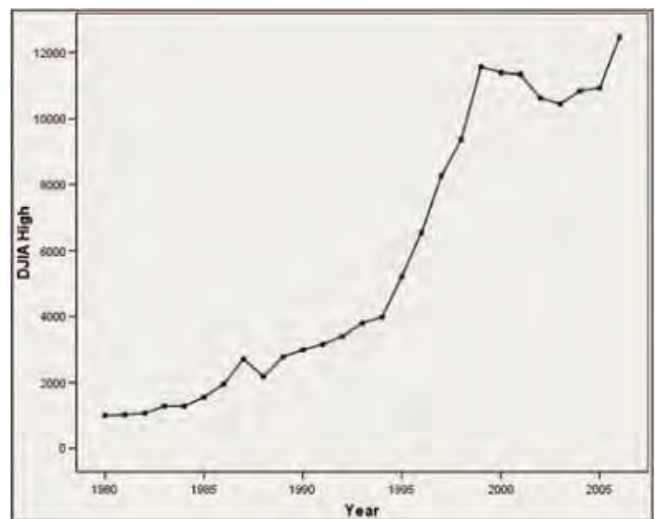
## Draw Time Series Graphs

If the value of a variable is measured at different points in time, the data are referred to as *time series data*.

A *time-series plot* is obtained by plotting the time in which a variable is measured on the horizontal axis and the corresponding value of the variable on the vertical axis. Line segments are then drawn connecting the points.

### Example

The accompanying SPSS-generated time-series graph shows the yearly high values of the Dow Jones Industrial Average (DJIA) for the N.Y. Stock Exchange. This graph shows a steady increase between the years 1980 and 2007, but the DJIA high values have not been so consistent in more recent years.



## Exercise Section 1.6 – Additional Displays

1. Identify the class width, class midpoints, and class boundaries for the given frequency distribution. Then construct the cumulative frequency distribution that corresponds to the frequency distribution.

a)

<i>Tar (mg) in Nonfiltered Cigarettes</i>	<i>Frequency</i>
10 – 13	1
14 – 17	0
18 – 21	15
22 – 25	7
26 – 29	2

b)

<i>Tar (mg) in Filtered Cigarettes</i>	<i>Frequency</i>
2 – 5	2
6 – 9	2
10 – 13	6
14 – 17	15

c)

<i>Weights (lb) of Discarded Metal</i>	<i>Frequency</i>
0.00 – 0.99	5
1.00 – 1.99	26
2.00 – 2.99	15
3.00 – 3.99	12
4.00 – 4.99	4

d)

<i>Weights (lb) of Discarded Plastic</i>	<i>Frequency</i>
0.00 – 0.99	14
1.00 – 1.99	20
2.00 – 2.99	1
3.00 – 3.99	4
4.00 – 4.99	2
5.00 – 5.99	1

2. Given listed amounts of Strontium-90 (in millibecquerels) in a simple random sample of baby teeth.

155 142 149 130 151 163 151 142 156 133 138 161 128 144 172  
 137 151 166 147 163 145 116 136 158 114 165 169 145 150 150  
 150 158 151 145 152 140 170 129 188 156

- Construct a dot plot of the amounts of Strontium-90. What does the dot plot suggest about the distribution of those amounts?
- Construct a stemplot of the amounts of Strontium-90. What does the stemplot suggest about the distribution of those amounts?
- Construct a frequency polygon of the amounts of Strontium-90. For the horizontal axis, use the midpoints of the class intervals in the frequency distribution: 110-119, 120-129, 130-139, ..., 180-189.
- Construct an ogive of the amounts of Strontium-90. For the horizontal axis, use the class boundaries corresponding to the class limits. How many of the amounts are below 150 millibecquerels?

3. Use the 62 weights if discarded plastic listed in Data set below

0.27 1.41 2.19 2.83 2.19 1.81 0.85 3.05 3.42 2.10 2.93 2.44 2.17 1.41 2.00  
 0.93 2.97 2.04 0.65 2.13 0.63 1.53 4.69 0.15 1.45 2.68 3.53 1.49 2.31 0.92  
 0.89 0.80 0.72 2.66 4.37 0.92 1.40 1.45 1.68 1.53 1.44 1.44 1.36 0.38 1.74  
 2.35 2.30 1.14 2.88 2.13 5.28 1.48 3.36 2.83 2.87 2.96 1.61 1.58 1.15 1.28  
 0.58 0.74

- Construct a dot plot of the weights of discarded plastic. What does the dot plot suggest about the distribution of the weights?
- Construct a stemplot of the weights of discarded plastic. What does the stemplot suggest about the distribution of the weights?
- Construct a frequency polygon of the weights of discarded plastic. For the horizontal axis, use the midpoints of the class intervals: 0.00-0.99, 1.00-1.99, 2.00-2.99, 3.00-3.99, 4.00-4.99, 5.00-5.99.
- Construct an ogive of the weights of discarded plastic. For the horizontal axis, use these classes boundaries:  $-0.005$ ,  $0.995$ ,  $1.995$ ,  $2.995$ ,  $3.995$ ,  $4.995$ ,  $5.995$ . How many of the weights are below 4 lb.?

4. In 1965, Intel cofounder Gordon Moore proposed what has since become known as Moore's law: the number of transistors per square inch on integrated circuits with double approximately every 18 months. The table below lists the number of transistors per square inch (in thousands) for several different years. Construct a time-series graph of the data.

Year	1971	1974	1978	1982	1985	1989	1993	1997	1999	2000	2002	2003
Transistors	2.3	5	29	120	275	1180	3100	7500	24,000	42,000	220,000	410,000

5. The following table shows the numbers of cell phone subscriptions (in thousands) in the U.S. for various years. Construct a time-series graph of the data. "Linear" growth would result in a graph that is approximately a straight line. Does the time-series graph appear to show linear growth?

Year	1985	1987	1989	1991	1993	1995	1997	1999	2001	2003	2005
Number	340	1231	3509	7557	16,009	33,786	55,312	86,047	128,375	158,722	207,900

6. The following table lists the marriage and divorce rates per 1000 people in the U.S. for selected years since 1900 (based on data from the Department of Health and Human Services). Construct a multiple bar graph of the data. Why do these data consist of marriage and divorce rates rather than total numbers of marriages and divorces? Comment on any trends that you observe in these rates, and give explanations for these trends.

Year	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
Marriage	9.3	10.3	12.0	9.2	12.1	11.1	8.5	10.6	10.6	9.8	8.3
Divorce	0.7	0.9	1.6	1.6	2.0	2.6	2.2	3.5	5.2	4.7	4.2

7. A car salesman records the number of cars he sold each week for the past year. The following frequency histogram shows the results

- What are the most frequent number of cars sold in a week?
- For how many weeks two cars sold?
- Determine the percentage of time two cars were sold.
- Describe the shape of the distribution





8. Use the data to create a stemplot

The midterm test scores for the seventh-period typing class are listed below

85 77 93 91 74 65 68 97 88 59 74 83 85 72 63 79

9. Use the data to create a stemplot. Twenty-four workers were surveyed about how long it takes them to travel to work each day. The data below are given in minutes

20 35 42 52 65 20 60 49 24 37 23 24

22 20 41 25 28 27 50 47 58 30 32 48

10. Find the original data from the stemplot

a)

<i>Stem</i>	<i>Leaves</i>
76	2 6 7
77	2 4 9
78	1 7

b)

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

c)

24	0 4 7
25	0 2 3 9 9
26	3 4 5 8 8 9
27	0 1 1 3 6 6
28	2 3 8

## Section 1.7 – Misrepresentations of Data

### Bad Graphs

Some graphs are bad in the sense that they contain errors. Some are bad because they are technically correct, but misleading. It is important to develop the ability to recognize bad graphs and identify exactly how they are misleading.

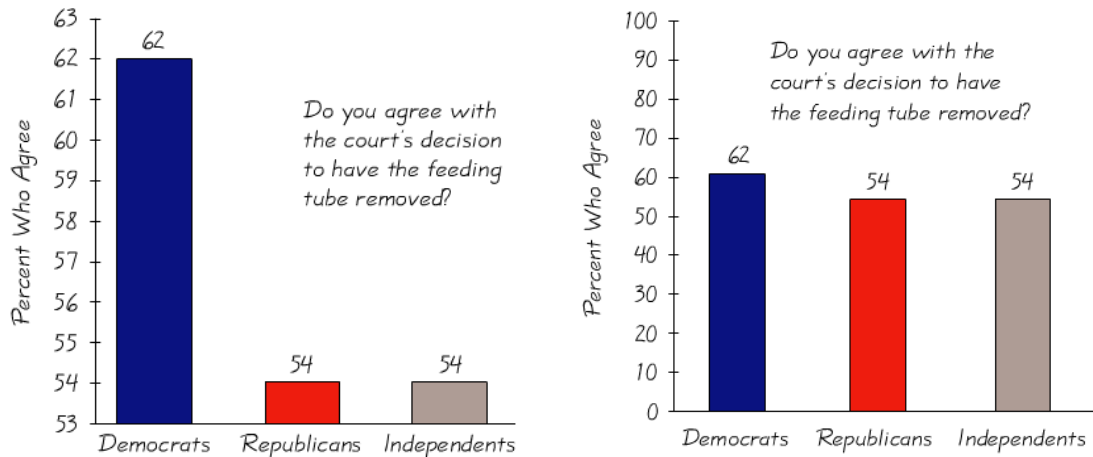
**Statistics:** The only science that enables different experts using the same figures to draw different conclusions. – **Evan Esar**

### Nonzero Axis

Some graphs are misleading because one or both of the axes begin at some value other than zero, so that differences are exaggerated.

### Example

The results of a CNN poll regarding the case of Schiavo is as shown in graph below



**Survey Results by Party**

This graph (on the left) creates the incorrect impression that significantly more Democrats agreed with the court's decision than Republicans or Independents. Since graph depicts the data objectively, it creates the more correct impression that the differences are not very substantial. Many people complained that it was deceptive, so CNN posted a modified graph similar to figure on the right.

### Pictographs

Drawings of objects, called pictographs, are often misleading. Three-dimensional objects - money bags, stacks of coins, army tanks (for army expenditures), people (for population sizes), barrels (for oil production), and houses (for home construction) are commonly used to depict data. These drawings can create false impressions that distort the data.

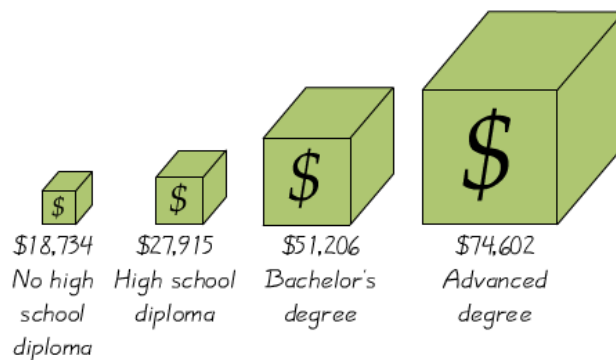
If you double each side of a square, the area does not merely double; it increases by a factor of four; if you double each side of a cube, the volume does not merely double; it increases by a factor of eight. Pictographs using areas and volumes can therefore be very misleading.

### Example



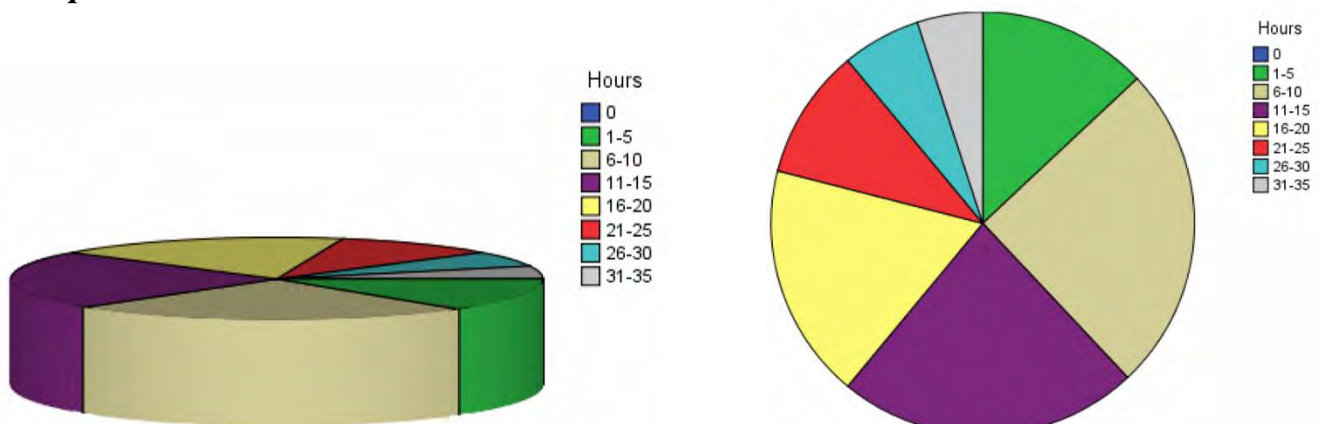
This picture is not misleading because the bars have same width, but it is somewhat too busy, and too difficult to understand.

### Example



It depicts one-dimensional data with three-dimensional boxes. Last box is 64 times as large as first box, but income is only 4 times as large.

### Example

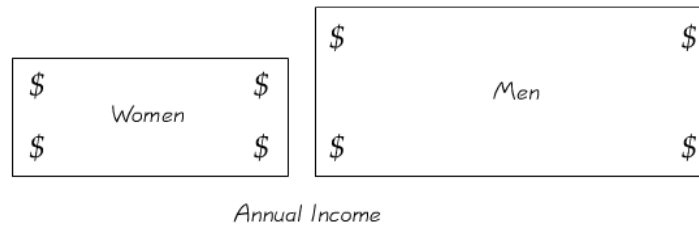


## **Guidelines for Constructing Good Graphics**

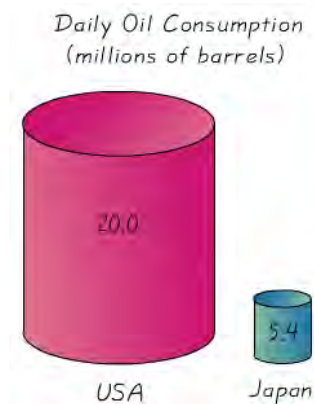
- ✓ Title and label the graphic axes clearly, providing explanations, if needed. Include units of measurement and a data source when appropriate.
- ✓ Avoid distortion. Never lie about the data.
- ✓ Minimize the amount of white space in the graph. Use the available space to let the data stand out. If scales are truncated, be sure to clearly indicate this to the reader.
- ✓ Avoid clutter, such as excessive gridlines and unnecessary backgrounds or pictures. Don't distract the reader.
- ✓ Avoid three dimensions. Three-dimensional charts may look nice, but they distract the reader and often lead to misinterpretation of the graphic.
- ✓ Do not use more than one design in the same graphic. Sometimes graphs use a different design in one portion of the graph to draw attention to that area. Don't try to force the reader to any specific part of the graph. Let the data speak for themselves.
- ✓ Avoid relative graphs that are devoid of data or scales.

## Exercise Section 1.7 – Misrepresentations of Data

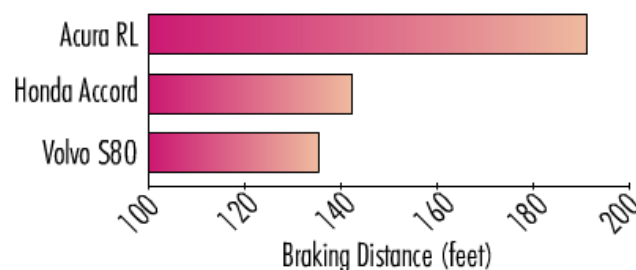
1. Assume that, as a newspaper reporter, you must graph data showing that increased smoking causes an increased risk of lung cancer. Given that people might be helped and lives might be saved by creating a graph that exaggerates the risk of lung cancer, is it ethical to construct such a graph?
2. The accompanying graph depicts average full-time incomes of women and men aged 18 and over. For a recent year, those incomes were \$37,197 for women and \$53,059 for men (based on data from the U.S. Census Bureau). Does the graph make a fair comparison of the data? Why or why not? If the graph distorts the data, construct a fair graph.



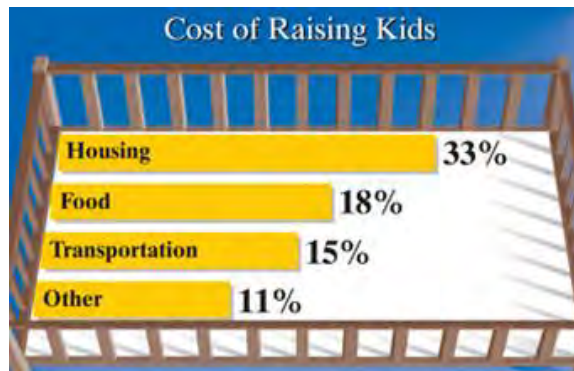
3. The accompanying graph uses cylinders to represent barrels of oil consumed by the U.S. and Japan. Does the graph distort the data or does it depict the data fairly? Why or why not? If the graph distorts the data, construct a graph that depicts the data fairly.



4. The accompanying graph shows the braking distances for different cars measured under the same conditions. Describe the ways in which this graph might be deceptive. How much greater is the braking distance of the Acura RL than the braking distance of the Volvo S80? Draw the graph in a way that depicts the data more fairly.



5. The graph represents the percentage of income a middle-income family will spend on their children



- How is the graphic misleading?
- What could be done to improve the graphics?

## Section 1.8 – Measures of Central Tendency

### Characteristics of center

Measures of center include mean and median, as tools for analyzing data. Not only determine the value of each measure of center, but also interpret those values.

### Definition

A *measure of center* is a value at the center or middle of a data set

### Mean

#### Definition

#### Definitions

The *arithmetic mean* of a variable is computed by adding all the values of the variable in the data set and dividing by the number of observations.

The *population arithmetic mean*,  $\mu$  (pronounced “mew”), is computed using all the individuals in a population. The population mean is a *parameter*.

The *sample arithmetic mean*,  $\bar{x}$  (pronounced “x-bar”), is computed using sample data. The sample mean is a *statistic*.

$$\text{mean} = \frac{\sum x}{N} \quad \begin{array}{l} \leftarrow \text{sum of all data values} \\ \leftarrow \text{number of data values} \end{array}$$

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{N} = \frac{\sum x}{N} \quad \text{is the mean of a set of *sample values*.}$$

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{N} = \frac{\sum x}{N} \quad \text{is the mean of all values in a *population*.}$$

### Notation

$\Sigma$  denotes the *sum* of a set of values.

$x$  is the *variable* usually used to represent the individual data values.

$n$  represents the *number of data values* in a *sample*. (*sample size*)

$N$  represents the *number of data values* in a *population*.

### Example

Find the mean of these first five word counts from men: 27,531; 15,684; 5,638; 27,997; and 25,433

### Solution

$$\begin{aligned}\bar{x} &= \frac{\sum x}{n} = \frac{27,531 + 15,684 + 5,638 + 27,997 + 25,433}{5} \\ &= \frac{102,283}{5} \\ &= 20,456.6\end{aligned}$$

The mean of the first five word counts is 20,456.6 words.

### Example

The following data represent the travel times (in minutes) to work for all seven employees of a start-up web development company.

23, 36, 23, 18, 5, 26, 43

- Compute the population mean of this data.
- Then take a simple random sample of  $n = 3$  employees. Compute the sample mean. Obtain a second simple random sample of  $n = 3$  employees. Again compute the sample mean.

### Solution

$$a) \mu = \frac{\sum x}{N} = \frac{23 + 36 + 23 + 18 + 5 + 26 + 43}{7} = \frac{174}{7} = 24.9 \text{ minutes}$$

- Obtain a simple random sample of size  $n = 3$  from the population of seven employees. Use this simple random sample to determine a sample mean. Find a second simple random sample and determine the sample mean.

1	2	3	4	5	6	7
23	36	23	18	5	26	43

```
123→rand      123
randInt(1,7)    5
                2
                6
```

$$\bar{x} = \frac{5 + 36 + 26}{3} = 22.3$$

```
789→rand      789
randInt(1,7)    2
                3
                6
```

$$\bar{x} = \frac{36 + 23 + 26}{3} = 28.3$$



## Median

### Definition

The **median** of a data set is the measure of center that is the **middle value** when the original data values are arranged in order of increasing (or decreasing) magnitude. The median is often denoted by  $\tilde{x}$  (*x-tilde*)

### Finding the Median

First **sort** the values (arrange them in order), the follow one of these

1. If the number of data values is odd, the median is the number located in the exact middle of the list.
2. If the number of data values is even, the median is found by computing the mean of the two middle numbers.

In order of **even** number of values: 5.40; 1.10; 0.42; 0.73; 0.48; 1.10

$$\begin{array}{cccccc} 0.42 & 0.48 & 0.73 & 1.10 & 1.10 & 5.40 \\ & & \uparrow & \uparrow & & \\ & & \text{Median} = \frac{0.73 + 1.10}{2} = \underline{0.195} \end{array}$$

In order of **odd** number of values: 5.40; 1.10; 0.42; 0.73; 0.48; 1.10; 0.66

$$\begin{array}{ccccccc} 0.42 & 0.48 & 0.66 & 0.73 & 1.10 & 1.10 & 5.40 \\ & & & \uparrow & & & \\ & & & \text{Median} = \underline{0.73} \end{array}$$

### Example

Find the median for this sample of data values: 27,531; 15,684; 5,638; 27,997; and 25,433

#### Solution

First sort the data: 5,638    15,684    25,433    27,531    27,997

Median is 25,433

### Example

Find the median for this sample of data values: 27,531, 15,684, 5,638, 27,997, 25,433 and 8,077

#### Solution

First sort the data: 5,638    8,077    15,684    25,433    27,531    27,997

Median is  $= \frac{15,684 + 25,433}{2} = \underline{20,558.5}$

### Example

The following data represent the travel times (in *minutes*) to work for all seven employees of a start-up web development company. 23, 36, 23, 18, 5, 26, 43

- Determine the median of this data.
- Suppose a new employee is hired who has a 130 minute commute. How does this impact the value of the mean and median?

### Solution

a) **Step 1:** 5, 18, 23, 23, 26, 36, 43

**Step 2:** There are  $n = 7$  observations.

**Step 2:**  $\frac{n+1}{2} = \frac{7+1}{2} = 4$

Median is 23

5, 18, 23, **23**, 26, 36, 43

b) Mean before new hire: 24.9 minutes

Median before new hire: 23 minutes

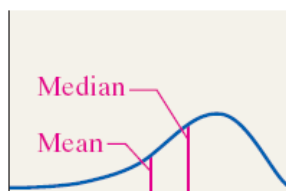
Mean after new hire: 38 minutes

Median after new hire: 24.5 minutes

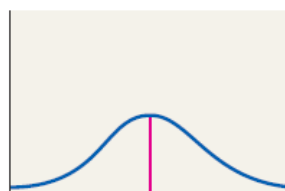
A numerical summary of data is said to be **resistant** if extreme values (very large or small) relative to the data do not affect its value substantially.

### Relation Between the Mean, Median, and Distribution Shape

<i>Distribution Shape</i>	<i>Mean vs. Median</i>
Skewed left	Mean substantially smaller than median
Symmetric	Mean Roughly equal to median
Skewed right	Mean substantially larger than median



(a) Skewed Left  
Mean < Median



(b) Symmetric  
Mean = Median



(c) Skewed Right  
Mean > Median

### Example

The following data represent the birth weights (in pounds) of 50 randomly sampled babies.

5.8	7.4	9.2	7.0	8.5	7.6	7.9	7.8	7.9	7.7	9.0	6.7	8.2	7.0
8.7	7.2	6.1	7.2	7.1	7.2	7.9	5.9	7.0	7.8	7.2	7.6	7.4	7.1
7.3	6.4	7.4	8.2	9.1	7.3	9.4	6.8	7.0	8.1	8.0	7.2	7.0	8.7
7.3	6.9	6.9	6.4	7.8	7.5	7.1	7.5						

- Find the mean and median.
- Describe the shape of the distribution
- Which measure of central tendency better describes the average birth weight?

### Solution

- Using the calculator:

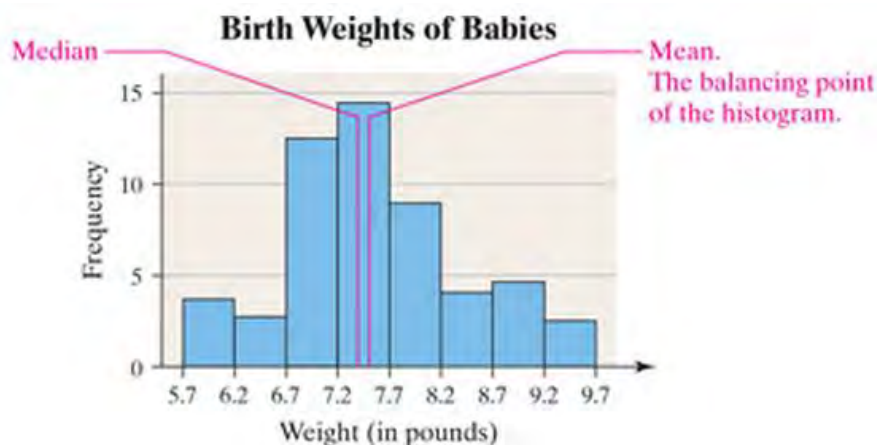
Mean:  $\bar{x} = 7.488 \approx 7.49$

Median  $M = 7.35$

```
1-Var Stats
x̄=7.488
Σx=374.4
Σx²=2835.1
Sx=.8029638973
σx=.7948937036
n=50
```

```
1-Var Stats
n=50
minX=5.8
Q1=7
Med=7.35
Q3=7.9
maxX=9.4
```

- 



The distribution is bell shaped.

- Because the mean and median are close in value, we use the mean as the measure of central tendency.

### ***Example***

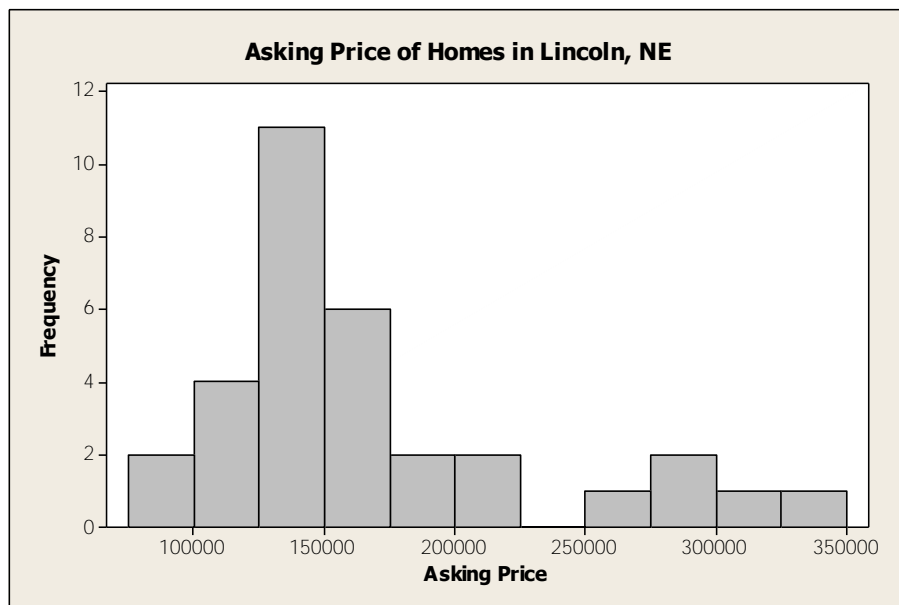
The following data represent the asking price of homes for sale in Lincoln, NE.

79,995	128,950	149,900	189,900
99,899	130,950	151,350	203,950
105,200	131,800	154,900	217,500
111,000	132,300	159,900	260,000
120,000	134,950	163,300	284,900
121,700	135,500	165,000	299,900
125,950	138,500	174,850	309,900
126,900	147,500	180,000	349,900

Find the mean and median. Use the mean and median to identify the shape of the distribution. Verify your result by drawing a histogram of the data.

### **Solution**

The mean asking price is \$168,320 and the median asking price is \$148,700. Therefore, we would conjecture that the distribution is skewed right.



## Mode

### Definition

The **mode** of a data set is the value that occurs with the greatest frequency

A data set can have one, more than one, or no mode

**Bimodal** two data values occur with the same greatest frequency

**Multimodal** more than two data values occur with the same greatest frequency

**No Mode** no data value is repeated

### Example

a) Find the mode of: 5.40 1.10 0.42 0.73 0.48 1.10  
**Mode** is 1.10

b) Find the mode of: 27 27 27 55 55 55 88 88 99  
**Mode** is 27 & 55 (**bimodal**)

c) Find the mode of: 1 2 3 4 5 6 7  
**No Mode**

## Midrange

### Definition

The midrange is the value midway between the maximum and minimum values in the original data set. It is found by adding the maximum data value to the minimum data value and then dividing the sum by 2:

$$\text{Midpoint} = \text{Midrange} = \frac{\text{minimum data value} + \text{maximum data value}}{2}$$

### Example

Find the midrange of these values: 27,531; 15,684; 5,638; 27,997; and 25,433

### Solution

$$\begin{aligned}\text{Midrange} &= \frac{\text{minimum data value} + \text{maximum data value}}{2} \\ &= \frac{5,638 + 27,997}{2} \\ &= 16,817.5\end{aligned}$$

## Critical Thinking

- Think about whether the results are reasonable.
- Think about the method used to collect the sample data.

### Example

For each of the following, identify a major reason why the mean and median are not meaningful statistics

- Zip codes: 1260, 77573, 77574, 90210, 77550
- Ranks of stress levels from different jobs: 2, 3, 1, 7, 9

### Solution

- The zip codes don't measure or count anything. The numbers are actually labels for geographic locations.
- The ranks reflect an ordering, but they don't measure or count anything. The rank of 1 might come from a job that has a stress level substantially greater than the stress level from the job with a rank of 2, so the different numbers don't correspond to the magnitude of the stress levels.

## Beyond the Basics of Measures of Center

### Mean from a Frequency Distribution

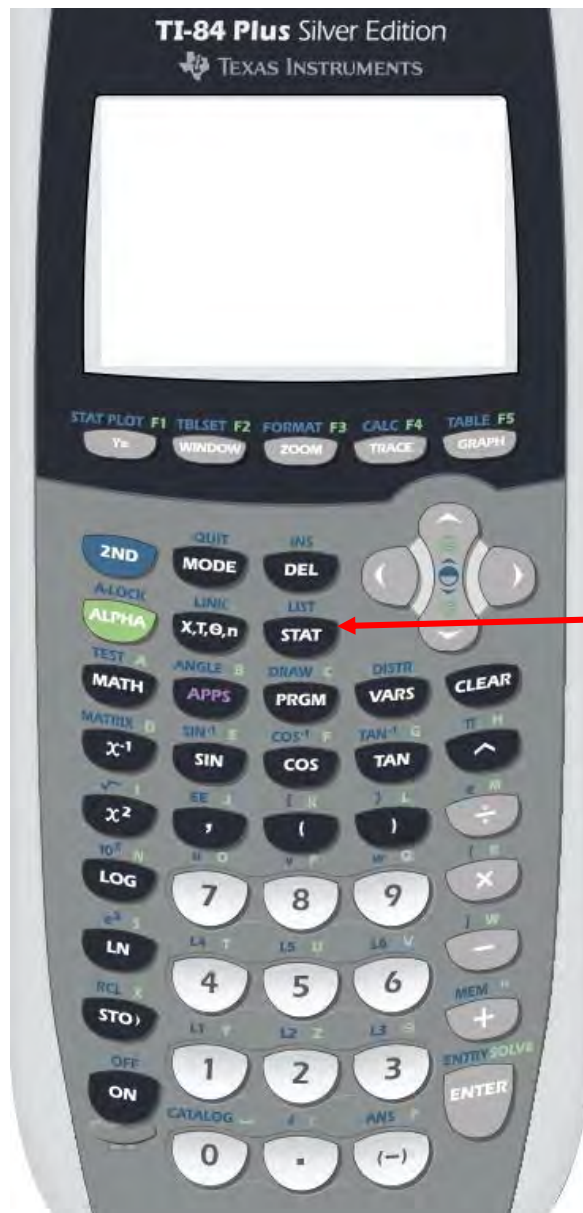
Assume that all sample values in each class are equal to the class midpoint.

*multiply each frequency and class midpoint, then add the products*

Mean from frequency distribution:  $\bar{x} = \frac{\sum (f \cdot x)}{\sum f}$  ← sum of frequencies

### Example

Word Counts	Frequency $f$	Class Midpoint $x$	$f \cdot x$
0 – 9,999	45	4,999.5	224,977.5
10,000 – 19,999	90	14,999.5	1,349,955
20,000 – 29,999	40	24,999.5	999,980
30,000 – 39,999	7	34,999.5	244,996.5
40,000 – 49,999	3	44,999.5	134,998.5
Totals:	$\Sigma f = 185$		$\Sigma f \cdot x = 2,954,907.5$
			$\bar{x} = \frac{\sum (f \cdot x)}{\sum f} = \frac{2,954,907}{185} = \underline{15,972.5}$



STAT

```

3000 CALC TESTS
1:Edit...
2:SortA(
3:SortD(
4:ClrList
5:SetUPEditor
  
```



L1	L2
4995.5	45.0
15000	90.0
25000	40.0
35000	7.0
45000	3.0
-----	-----
L2(0)=45	



Click on



→ CALC → 1

```

EDIT CALC TESTS
1:1-Var Stats
  
```



```

1-Var Stats
x=15972.5
Σx=2954907.5
Σx²=6.1E10
Sx=8668.1
σx=8644.6
n=185.0
  
```

Mean Value



## Weighted Mean

When data values are assigned different weights, we can compute a **weighted mean**.

$$\text{weighted mean: } \bar{x} = \frac{\sum (w \cdot x)}{\sum w}$$

### Example

In her first semester of college, a student of the author took five courses. Her final grades along with the number of credits for each course were: *A* (3 credits); *A* (4 credits); *B* (3 credits); *C* (3 credits) and *F* (1 credit). The grading system assigns quality points to letter grades as follows:

$$A = 4; B = 3; C = 2; D = 1; F = 0.$$

Compute her grade point average.

### Solution

Weights = number of credits:  $w = 3, 4, 3, 3, 1$ .

Replace A, B, C, D, and F with the corresponding quality points:  $x = 4, 4, 3, 2, 0$ .

$$\begin{aligned}\bar{x} &= \frac{\sum (w \cdot x)}{\sum w} \\ &= \frac{(3 \times 4) + (4 \times 4) + (3 \times 3) + (3 \times 2) + (1 \times 0)}{3 + 4 + 3 + 3 + 1} \\ &= \frac{43}{14} \\ &= 3.07\end{aligned}$$

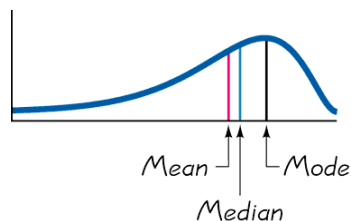
## Skewed and Symmetric

### Definition

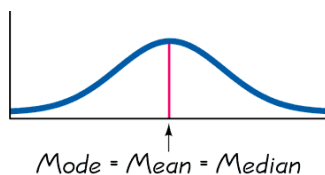
A distribution of data is **skewed** if it is not symmetric and extends more to one side than the other.

A distribution of data is **symmetric** if the left half of its histogram is roughly a mirror image of its right half.

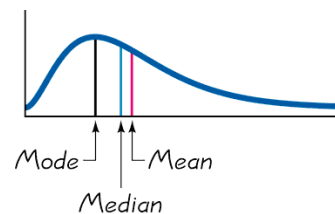
The mean and median cannot always be used to identify the shape of the distribution.



Skewed to the Left (Negatively)

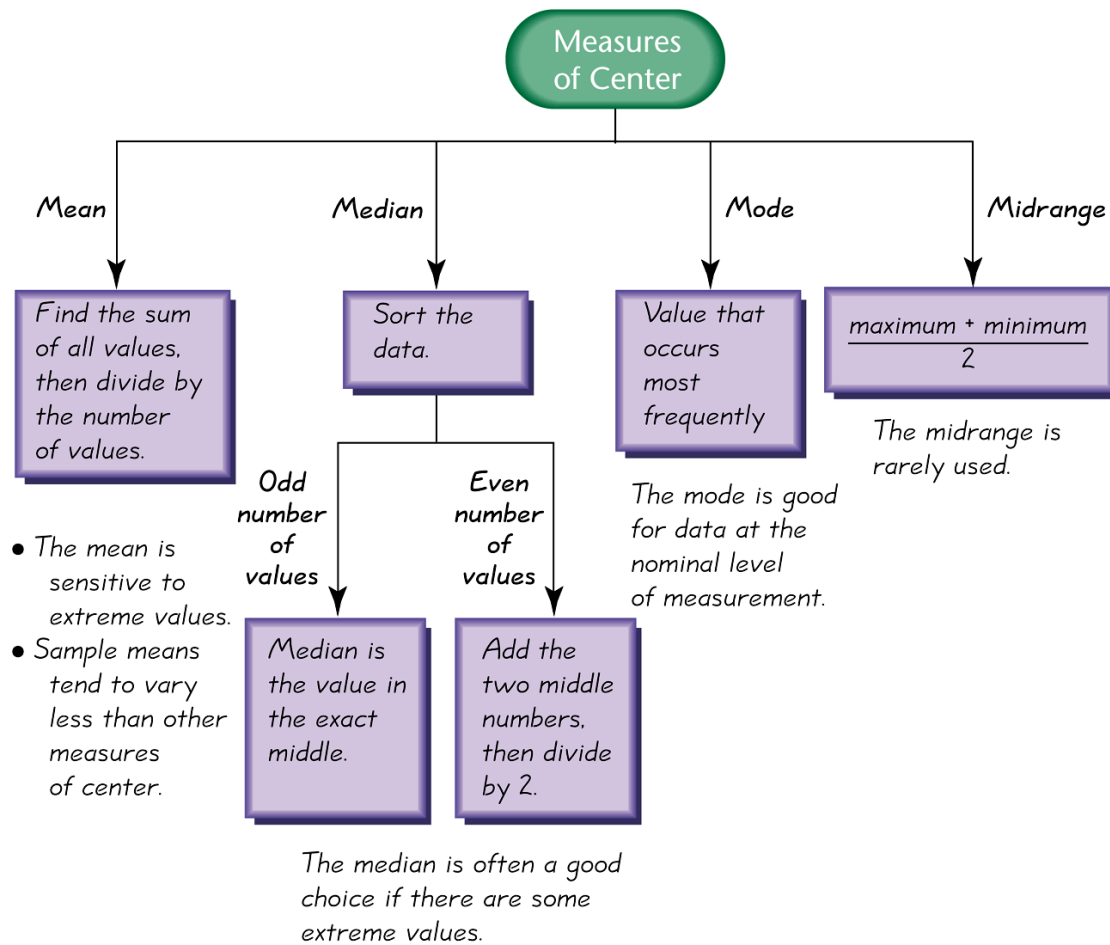


Symmetric



Skewed to the Right (Positively)





## Exercises      Section 1.8 – Measures of Central Tendency

1. In what sense are the mean, median, mode and midrange measures of “center”?
2. A headline in USA Today stated that “Average family income drops 2.3%.” What is the role of the term average in statistics? Should another term be used in place of average?
3. In an editorial, the Poughkeepsie Journal printed this statement: “The median price – the price exactly in between the highest and lowest -- ...” Does that statement correctly describes the median? Why or why not?
4. A simple random sample of pages from Merriam-Webster’s Collegiate Dictionary, 11th edition, was obtained. Listed below are the numbers of words defined on those pages. Given that this dictionary has 1459 pages with defined words, estimate the total number of defined words in the dictionary.  
51   63   36   43   34   62   73   39   53   79  
Find the   a) mean   b) median   c) mode   d) midrange  
e) Is that estimate likely to be an accurate estimate of the number of words in the English language?
5. The National Highway Traffic Administration conducted crash tests of child booster seats for cars. Listed below are results from those tests, with the measurements given in hic (standard head injury condition units).  
774   249   1210   546   431   612  
Find the   a) mean   b) median   c) mode   d) midrange  
e) According to the safety requirement, the hic measurement should be less than 1000 hic. Do the results suggest that all of the child booster seats meet the specified requirement?
6. The insurance Institution for Highway Safety conducted tests with crashes of new cars traveling at 6 mi/h. The total cost of the damages was found for a simple random sample of the tested cars and listed below  
\$7448   \$4911   \$9051   \$6374   \$4277  
Find the   a) mean   b) median   c) mode   d) midrange  
e) Do the different measures of center differ very much?
7. Listed below are the durations (in hours) of a simple random sample of all flights (as of this writing) of NASA’s Space Transport System (space shuttle).  
73   95   235   192   165   262   191   376   259   235   381   331   221   244   0  
Find the   a) mean   b) median   c) mode   d) midrange  
e) How might that duration time be explained?

8. Listed below are the playing times (in seconds) of songs that were popular at the time of this writing.

448 242 231 246 246 293 280 227 213 262 239 213 258 255 257

Find the a) mean b) median c) mode d) midrange

e) Is there on time that is very different from the others?

9. Listed below are numbers of satellites in orbit from different countries.

158 17 15 17 7 3 5 1 8 3 4 2 4 1 2 3 1 1 1 1 1 1 1 1

Find the a) mean b) median c) mode d) midrange

e) Does on country have an exceptional number of satellites?

f) Can you guess which country has the most satellites?

10. Listed below are costs (in dollars) of roundtrip flights from JFK airport in NY City to San Francisco. (All flights involve one stop and a two-week stay.) The airlines are US Air, Continental, Delta, United, American, Alaska, and Northwest.

30 Days in Advance	244	260	264	264	278	318	280
1 Day in Advance	456	614	567	943	628	1088	536

a) Find the mean and median for each then compare the two sets of results.

b) Does it make much difference if the tickets are purchased 30 days in advance or 1 day in advance?

11. The trend of thinner Miss America winners has generated charges that the contest encourages unhealthy diet habits among young women. Listed below are body mass indexes (BMI) for Miss America winners from two different periods.

BMI (1920 – 1930)	20.4	21.9	22.1	22.3	20.3	18.8	18.9	19.4	18.4	19.1
BMI – (from recent winners)	19.5	20.3	19.6	20.2	17.8	17.9	19.1	18.8	17.6	16.8

Find the mean and median for each then compare the two sets of results.

12. Find the mean of the data summarized in the given frequency distribution.

a)

<i>Tar (mg) in Nonfiltered Cigarettes</i>	<i>Frequency</i>
10 – 13	1
14 – 17	0
18 – 21	15
22 – 25	7
26 – 29	2

b)

<i>Pulse Rates of Females</i>	<i>Frequency</i>
60 – 69	12
70 – 79	14
80 – 89	11
90 – 99	1
100 – 109	1
110 – 119	0
120 – 129	1

13. A student of the author earned grades of B, C, B, A, and D. Those courses has these corresponding numbers credit hours: 3, 3, 4, 4, and 1. The grading system assigns quality points to letter grades as follows: A = 4; B = 3; C = 2; D = 1; F = 0. Compute the grade point average (GPA) and round the result with two decimal places. If the Dean's list requires a GPA 3.00 or greater, did this student make the Dean's list?
14. A student of the author earned grades of 92, 83, 77, 84, and 82 on her five regular tests. She earned grades of 88 on the final exam and 95 on her class projects. Her combined homework grade was 77. The five regular tests count for 60% of the final grade, the final exam counts for 10%, the project counts for 15%, and homework counts for 15%. What is her weighted mean grade? What letter grade did she earn? (A, B, C, D, or F)
15. You are taking a class in which your grade is determined from five sources: 50% from your test mean, 15% from your midterm, 20% from your final exam, 10% from your computer lab work, and 5% from your homework. Your scores are 86 (test mean), 96 (midterm), 82 (final exam), 98 (computer lab), and 100 (homework). What is the weighted mean of your scores? If the minimum average for an A is 90, did you get an A?
16. During a quality assurance check, the actual coffee contents (in ounces) of six jars of instant coffee were recorded as 6.03, 5.59, 6.40, 6.00, 5.99, and 6.02.
- Find the mean and the median of the coffee content.
  - The third value was incorrectly measured and is actually 6.04. Find the mean and median of the coffee content again.
  - Which measure of central tendency, the mean or the median, was affected more by the data entry error?
17. The table below shows the U.S. exports (in billions of dollars) to 19 countries for a recent year.

<b><i>U.S. Exports</i></b> (in billions of dollars)		
Canada: 261.1	Mexico: 151.2	Germany: 54.5
Taiwan: 24.9	Netherlands: 39.7	China: 69.7
Australia: 22.2	Malaysia: 12.9	Switzerland: 22.0
Saudi Arabia: 12.5	United Kingdom: 53.6	Japan: 65.1
South Korea: 34.7	Singapore: 27.9	France: 28.8
Brazil: 32.3	Belgium: 28.9	Italy: 15.5
Thailand: 9.1		

- Find the mean and the median.
- Find the mean and median without the U.S. exports to Canada. Which measure of central tendency, the mean or the median, was affected more by the elimination of the Canadian exports?
- The U.S. Exports to India were \$17.7 billion. Find the mean and median with the Indian exports added to the original data set. Which measure of central tendency was affected more by adding the Indian exports?

## Section 1.9 – Measures of Dispersion

### Basic Concepts of Variation

#### Range

##### *Definition*

The **range** of a set of data values is the difference between the maximum data value and the minimum data value.

$$\text{Range} = (\text{maximum data value}) - (\text{minimum data value})$$

*It is very sensitive to extreme values; therefore not as useful as other measures of variation.*

##### *Example*

India has 1 satellite used for military and intelligence purposes, Japan has 3, and Russia has 14. Find the range of the sample values of 1, 3, and 14.

##### *Solution*

$$\begin{aligned}\text{Range} &= (\text{maximum data value}) - (\text{minimum data value}) \\ &= 14 - 1 \\ &= 13.0\end{aligned}$$

##### *Definition*

The **standard deviation** of a set of sample values, denoted by  $s$ , is a measure of variation of values about the mean. It is a type of average deviation of values from the mean that is calculated.

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \quad \text{Sample standard deviation}$$

$$s = \sqrt{\frac{n \left( \sum x^2 \right) - \left( \sum x \right)^2}{n(n - 1)}} \quad \text{Sample standard deviation}$$

### Standard Deviation - Important Properties

- 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.

## Example

Find the standard deviation of the numbers: 7, 9, 18, 22, 27, 29, 32, 40.

### Solution

$$s = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n-1}}$$
$$= \sqrt{\frac{5132 - 8(23)^2}{8-1}}$$
$$\approx 11.34$$

```
1-Var Stats
x̄=23.000
Σx=184.000
Σx²=5132.000
Sx=11.339
σx=10.607
n=8.000
```

## Standard Deviation of a Population

The standard deviation  $\sigma$  (lowercase sigma) of a population is given by the formula

Population standard deviation

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n-1}}$$

## Variance of a Sample and a Population

### Definition

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

Sample variance:  $s^2$  square of the standard deviation  $s$ .

Population variance:  $\sigma^2$  square of the population standard deviation  $\sigma$ .

### Notation

$s$  = sample standard deviation

$s^2$  = sample variance

$\sigma$  = population standard deviation

$\sigma^2$  = population variance

Population standard deviation

```
1-Var Stats
x̄=79
Σx=790
Σx²=63374
Sx=10.3494498
σx=9.818350167
n=10
```

Sample standard deviation

```
1-Var Stats
x̄=80
Σx=320
Σx²=25794
Sx=8.041558721
σx=6.964194139
n=4
```

## Unbiased Estimator

The sample variance  $s^2$  is an **unbiased estimator** of the population variance  $\sigma^2$ , which means values of  $s^2$  tend to target the value of  $\sigma^2$  instead of systematically tending to overestimate or underestimate  $\sigma^2$ .

## Using and Understanding Standard Deviation

**Range Rule of Thumb** is based on the principle that for many data sets, the vast majority (such as 95%) of sample values lie within two standard deviations of the mean

### *Interpreting a Known Value of the Standard Deviation*

Informally define **usual** values in a data set to be those that are typical and not too extreme. Find rough estimates of the minimum and maximum “usual” sample values as follows:

$$\text{Minimum “usual” value} = (\text{mean}) - 2 \times (\text{standard deviation})$$

$$\text{Maximum “usual” value} = (\text{mean}) + 2 \times (\text{standard deviation})$$

### Estimating a Value of the Standard Deviation $s$

To roughly estimate the standard deviation from a collection of known sample data use

$$s \approx \frac{\text{range}}{4}$$

Where range = (maximum value) – (minimum value)

### *Example*

The Wechsler Adult intelligence Scale involves an IQ test designed so that the mean score is 100 and the standard deviation is 15. Use the range rule thumb to find the minimum and maximum “usual” IQ scores. Then determine whether an IQ score of 135 would be considered “unusual”

### **Solution**

$$\text{Mean} = 100$$

$$\text{Standard deviation} = 15$$

$$\begin{aligned}\text{Minimum “usual” value} &= (\text{mean}) - 2 \times (\text{standard deviation}) \\ &= 100 - 2(15) \\ &= 70\end{aligned}$$

$$\begin{aligned}\text{Maximum “usual” value} &= (\text{mean}) + 2 \times (\text{standard deviation}) \\ &= 100 + 2(15) \\ &= 130\end{aligned}$$

### Example

Use the range of thumb to estimate the standard deviation of the sample of 100 FICO credit rating scores listed in the table below.

708	713	781	809	797	793	711	681	768	611	698	836	768
532	657	559	741	792	701	753	745	681	598	693	743	444
502	739	755	835	714	517	787	714	497	636	637	797	568
714	618	830	579	818	654	617	849	798	751	731	850	591
802	756	689	789	628	692	779	756	782	760	503	784	591
834	694	795	660	651	696	638	635	795	519	682	824	603
709	777	829	744	752	783	630	753	661	604	729	722	706
594	664	782	579	796	611	709	697	732				

### Solution

Those scores have a minimum of 444 and a maximum of 850.

$$\begin{aligned}s &\approx \frac{\text{range}}{4} \\&= \frac{850 - 444}{4} \\&= 101.5\end{aligned}$$

### Properties of the Standard Deviation

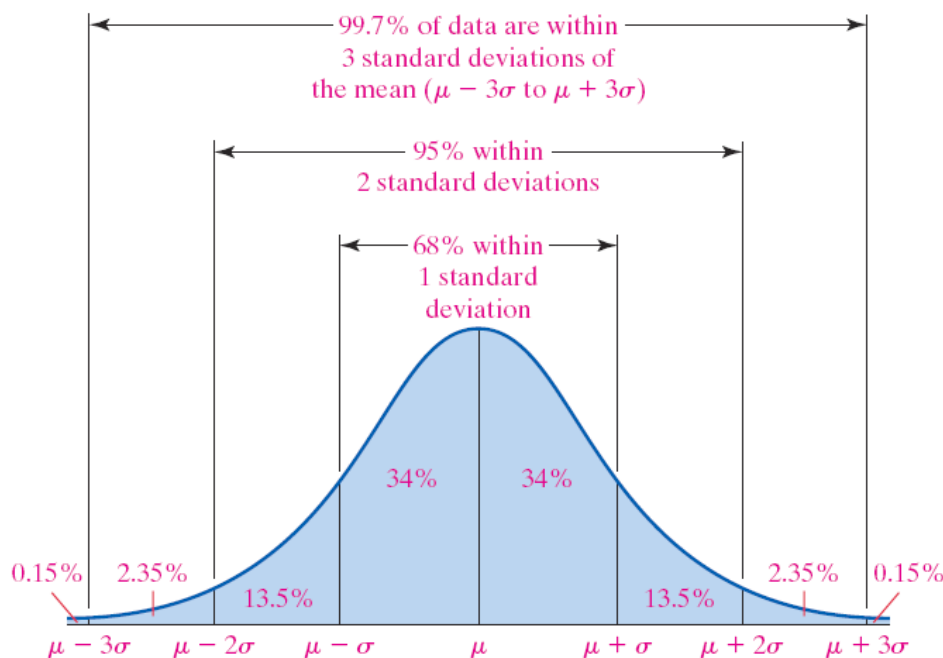
- ✓ The standard deviation measures the **variation** among data values
- ✓ Values close together have a small standard deviation, but values with much more variation have a larger standard deviation
- ✓ Has the same units of measurement as the original data
- ✓ For many data sets, a value is *unusual* if it differs from the mean by more than two standard deviations
- ✓ When comparing variation in two different data sets, compare the standard deviation only if they use the same scale and units, and they have means that are approximately the same.



## Empirical (or 68-95-99.7) Rule

Another concept that is helpful in interpreting the value of a standard deviation is the *empirical 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.



### Example

Empirical Rule IQ scores have a bell-shaped distribution with mean of 100 and a standard deviation of 15. What percentages of IQ scores are between 70 and 130?

### Solution

$$130 = 100 + 15 + 15$$

$$70 = 100 - 15 - 15$$

70 and 130 are each exactly 2 standard deviation away from the mean 100.

$$2 \text{ standard deviation} = 2s = 2(15) = 30$$

Therefore, 2 standard deviation from the mean is

$$100 - 30 = 70$$

$$100 + 30 = 130$$

The empirical rule tells us that about 95% of all values are within 2 standard deviation of the mean, so about 95% of all IQ scores are between 70 and 130.

## Chebyshev's Theorem

The proportion (or fraction) of any set of data lying within  $K$  standard deviations of the mean is always at least  $\left(1 - \frac{1}{K^2}\right) 100\%$ , where  $K$  is any positive number greater than 1.

- For  $K = 2$ , at least  $3/4$  (or 75%) of all values lie within 2 standard deviations of the mean.
- For  $K = 3$ , at least  $8/9$  (or 89%) of all values lie within 3 standard deviations of the mean.

## Chebyshev's Inequality

For any data set or distribution, at least  $\left(1 - \frac{1}{K^2}\right) 100\%$  of the observations lie within  $k$  standard deviations of the mean, where  $k$  is any number greater than 1. That is, at least  $\left(1 - \frac{1}{K^2}\right) 100\%$  of the data lie between  $\mu - k\sigma$  and  $\mu + k\sigma$  for  $k > 1$ .

**Note:** We can also use Chebyshev's Inequality based on sample data.

### Example

Chebyshev's Theorem IQ scores have a mean of 100 and a standard deviation of 15. What can we conclude from Chebyshev's Theorem?

### Solution

We can conclude that:

At least  $\frac{3}{4}$  (or 75%) of IQ scores are within 2 standard deviation of the mean (between 70 and 130).

At least  $\frac{8}{9}$  (or 89%) of IQ scores are within 3 standard deviation of the mean (between 55 and 145).

## Standard Deviation Defined

For a particular data value of  $x$ , the amount of deviation is  $x - \bar{x}$ . Those deviations could be a negative numbers, and the sum could be zero. To get statistic that measures variation (instead of always zero), we need to avoid canceling out of negative and positive numbers. We can get the **mean absolute deviation** (or **MAD**), which is the mean distance of the data from the mean:

$$\text{mean absolute deviation} = \frac{\sum |x - \bar{x}|}{n}$$

### Example

The following data represent the serum HDL cholesterol of the 54 female patients of a family doctor.

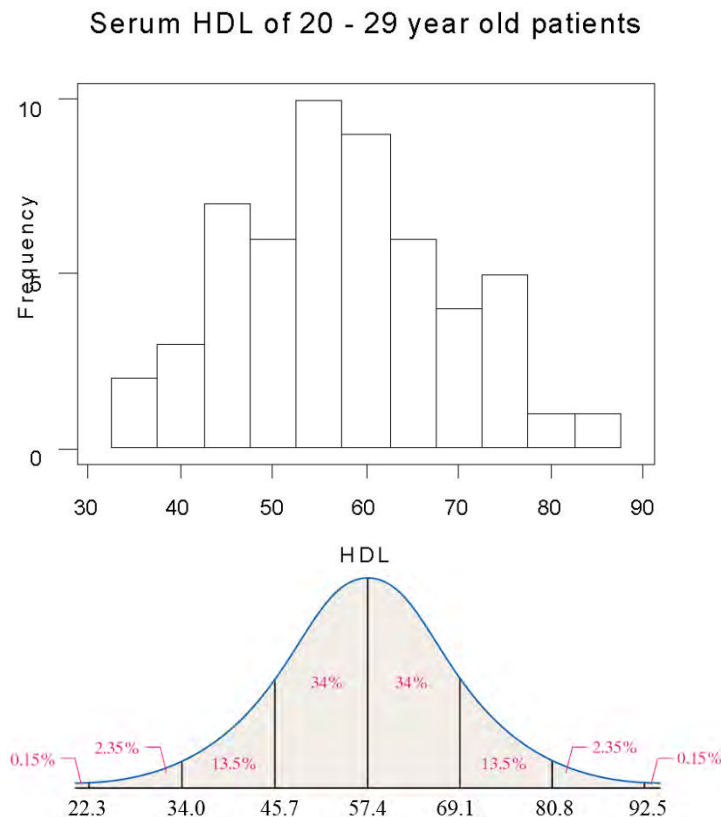
41	48	43	38	35	37	44	44	44
62	75	77	58	82	39	85	55	54
67	69	69	70	65	72	74	74	74
60	60	60	61	62	63	64	64	64
54	54	55	56	56	56	57	58	59
45	47	47	48	48	50	52	52	53

- Compute the population mean and standard deviation.
- Draw a histogram to verify the data is bell-shaped.
- Determine the percentage of all patients that have serum HDL within 3 standard deviations of the mean according to the Empirical Rule.
- Determine the percentage of all patients that have serum HDL between 34 and 69.1 according to the Empirical Rule.
- Determine the actual percentage of patients that have serum HDL between 34 and 69.1.
- Determine the percentage of patients that have serum HDL within 3 standard deviations of the mean.
- Determine the actual percentage of patients that have serum HDL between 34 and 80.8 (within 3 SD of mean).

### Solution

- Using a TI-83 plus graphing calculator, we find  $\mu = 57.4$  and  $\sigma = 11.7$

b)



- c) According to the Empirical Rule, 99.7% of the all patients that have serum HDL within 3 standard deviations of the mean.
- d)  $13.5\% + 34\% + 34\% = 81.5\%$  of all patients will have a serum HDL between 34.0 and 69.1 according to the Empirical Rule.
- e) 45 out of the 54 or 83.3% of the patients have a serum HDL between 34.0 and 69.1.
- f)  $\left(1 - \frac{1}{3^2}\right) 100\% = 88.9\%$
- g)  $\frac{52}{54} \approx 0.96 = 96\%$

## Definition

The **coefficient of variation** (or **CV**) for a set of nonnegative sample or population data, expressed as a percent, describes the standard deviation relative to the mean.

$$\text{Sample} \\ CV = \frac{s}{\bar{x}} \cdot 100\%$$

$$\text{Population} \\ CV = \frac{\sigma}{\mu} \cdot 100\%$$

## Example

Compare the variation in heights of men to the variation in weights of men, using these sample results obtained from data below

### Men heights

70.8	66.2	71.7	68.7	67.6	69.2	66.5	67.2	68.3	65.6	63.0	68.3	73.1	67.6
68.0	71.0	61.3	76.2	66.3	69.7	65.4	70.0	62.9	68.5	68.3	69.4	69.2	68.0
71.9	66.1	72.4	73.0	68.0	68.7	70.3	63.7	71.1	65.6	68.3	66.3		

### Men weights

169.1	144.2	179.3	175.8	152.6	166.8	135.0	201.5	175.2	139.0	156.3	186.6	191.1
151.3	209.4	237.1	176.7	220.6	166.1	137.4	164.2	162.4	151.8	144.1	204.6	193.8
172.9	161.9	174.8	169.8	213.3	198.0	173.3	214.5	137.1	119.5	189.1	164.7	170.1
151.0												

## Solution

The heights yield:  $\bar{x} = 68.34$  in. and  $s = 3.02$  in.

The weights yield:  $\bar{x} = 172.55$  lb. and  $s = 26.33$  lb.

```
1-Var Stats
x̄=68.335
Σx=2733.400
Σx²=187142.480
Sx=3.020
σx=2.982
↓n=40.000
```

### *Heights*

$$\text{Heights} \quad CV = \frac{s}{\bar{x}} \cdot 100\% = \frac{3.02}{68.34} \cdot 100\% = \underline{4.42\%}$$

$$\text{Weights} \quad CV = \frac{s}{\bar{x}} \cdot 100\% = \frac{26.33}{172.55} \cdot 100\% = \underline{15.26\%}$$

We can see that heights (with  $CV = 4.42\%$ ) have considerably less variation than weights (with  $CV = 15.26\%$ ). This makes intuitive sense, because the weights among men vary much more than heights. It is very rare to see two adult men with one of them being twice as tall as the other, but it is much more common to see two men with one of them weighing twice as much as the other.

## Exercises      Section 1.9 – Measures of Dispersion

1. In statistics, how do variation and variance differ?
2. Collegiate Dictionary has 1459 pages of defined words. Listed below are the numbers of defined words per page for a simple random sample of those pages. If we use this sample as a basis for estimating the total number of defined words in the dictionary.  
51   63   36   43   34   62   73   39   53   79
  - a) Find the range, variance, and standard deviation.
  - b) How does the variation of these numbers affect our confidence on the accuracy of the estimate?
3. The National Highway Traffic Administration conducted crash tests of child booster seats for cars. Listed below are results from those tests, with the measurements given in hic (standard head injury condition units).  
774   249   1210   546   431   612
  - a) Find the range, variance, and standard deviation
  - b) According to the safety requirement, the hic measurement should be less than 1000 hic. Do the results suggest that all of the child booster seats meet the specified requirement?
4. The insurance Institution for Highway Safety conducted tests with crashes of new cars traveling at 6 mi/h. The total cost of the damages was found for a simple random sample of the tested cars and listed below  
\$7448   \$4911   \$9051   \$6374   \$4277
  - a) Find the range, variance, and standard deviation
  - b) Do the different measures of center differ very much?
5. Listed below are the durations (in hours) of a simple random sample of all flights (as of this writing) of NASA's Space Transport System (space shuttle).  
73   95   235   192   165   262   191   376   259   235   381   331   221   244   0
  - a) Find the range, variance, and standard deviation
  - b) How might that duration time be explained?
6. Listed below are the playing times (in seconds) of songs that were popular at the time of this writing.  
448   242   231   246   246   293   280   227   213   262   239   213   258   255   257
  - a) Find the range, variance, and standard deviation
  - b) Is there on time that is very different from the others?
7. Listed below are numbers of satellites in orbit from different countries.  
158   17   15   17   7   3   5   1   8   3   4   2   4   1   2   3   1   1   1   1   1   1   1
  - a) Find the range, variance, and standard deviation
  - b) Does on country have an exceptional number of satellites?

8. Listed below are costs (in dollars) of roundtrip flights from JFK airport in NY City to San Francisco. (All flights involve one stop and a two-week stay.) The airlines are US Air, Continental, Delta, United, American, Alaska, and Northwest.

30 Days in Advance	244	260	264	264	278	318	280
1 Day in Advance	456	614	567	943	628	1088	536

Find the coefficient of variation for each of the two sets of data, then compare the variation.

9. The trend of thinner Miss America winners has generated charges that the contest encourages unhealthy diet habits among young women. Listed below are body mass indexes (BMI) for Miss America winners from two different periods.

BMI (1920 – 1930)	20.4	21.9	22.1	22.3	20.3	18.8	18.9	19.4	18.4	19.1
BMI – (from recent winners)	19.5	20.3	19.6	20.2	17.8	17.9	19.1	18.8	17.6	16.8

Find the coefficient of variation for each of the two sets of data, then compare the variation.

10. Find the Standard Deviation from the frequency distribution and find the standard deviation of sample summarized in a frequency distribution table by using the formula

$$s = \sqrt{\frac{n \left[ \sum (f \cdot x^2) \right] - \left[ \sum (f \cdot x) \right]^2}{n(n-1)}}, \text{ where } x \text{ represents the class midpoint, } f \text{ represents the class frequency, and } n \text{ represents the total number of sample values.}$$

a)

<b><i>Tar (mg) in Nonfiltered Cigarettes</i></b>	<b><i>Frequency</i></b>
10 – 13	1
14 – 17	0
18 – 21	15
22 – 25	7
26 – 29	2

b)

<b><i>Pulse Rates of Females</i></b>	<b><i>Frequency</i></b>
60 – 69	12
70 – 79	14
80 – 89	11
90 – 99	1
100 – 109	1
110 – 119	0
120 – 129	1

11. Heights of women have a bell-shaped distribution with a mean of 161 cm and a standard deviation of 7 cm. Using the empirical rule, what is the approximate percentage of women between
- 154 cm and 168 cm?
  - 147 cm and 175 cm?
12. The author's Generac generator produces voltage amounts with a mean of 125.0 volts and standard deviation of 0.3 volts, and the voltages have a bell-shaped distribution. Using the empirical rule, what is the approximate percentage of voltage amounts between
- 124.4 volts and 125.6 volts?
  - 124.1 volts and 125.9 volts?

13. The mean value of land and buildings per acre from a sample of farms is \$1,500, with a standard deviation of \$200. Using the empirical rule, estimate the percent of farms whose land and building values per acre are between \$1,300 and \$1,700. (Assume the data set has a bell-shaped distribution.)
14. The mean value of land and buildings per acre from a sample of farms is \$2,400, with a standard deviation of \$450. Using the empirical rule, between what two values do about 95% of the data lie? (Assume the data set has a bell-shaped distribution.)
15. Heights of women have a bell-shaped distribution with a mean of 161 cm and a standard deviation of 7 cm. Using Chebyshev's Theorem, what do we know about the percentage of women with heights that are within 2 standard deviations of the mean? What are the minimum and maximum heights that are within 2 standard deviations of the mean?
16. The author's Generac generator produces voltage amounts with a mean of 125.0 volts and standard deviation of 0.3 volts. Using Chebyshev's Theorem, what do we know about the percentage of voltage amounts that are within 3 standard deviations of the mean? What are the minimum and maximum voltage amounts that are within 3 standard deviations of the mean?
17. The mean time in a women's 400-meter dash is 57.07 seconds, with a standard deviation of 1.05 seconds. Apply Chebyshev's Theorem to the data using  $k = 2$ . Interpret the results.
18. The number of gallons of water consumed per day by a small village are listed. Make a frequency distribution (using five classes) for the data set. Then approximate the population mean and the population standard deviation of the data set.  
 167 180 192 173 145 151 174 175 178 160  
 195 224 244 146 162 146 177 163 149 188
19. To get the best deal on a microwave oven, Jeremy called six appliance stores and asked the cost of a specific model. The prices he was quoted are listed below:  
 \$325 \$384 \$156 \$210 \$219 \$284  
 Find the variance for the given data.
20. Compare the variation in heights to the variation in weights of thirteen-year old girls. The heights (in inches) and weights (in pounds) of nine randomly selected thirteen-year old girls as listed below  
 Heights (inches): 59.3 61.2 62.6 64.7 60.1 58.3 64.6 63.7 66.1  
 Weights (pounds): 87 96 91 119 96 90 123 98 139  
 Find the coefficient of variation for each of the two sets of data, then compare the variation
21. The amount of Jen's monthly phone bill is normally distributed with a mean of \$56 and a standard deviation of \$9. What percentage of her phone bills are between \$29 and \$83? Use the empirical rule to solve.



## Section 1.10 – Measures of Position, Outliers, and Boxplots

This section introduces measures of relative standing, which numbers are showing the location of data values relative to the other values within a data set. They can be used to compare values from different data sets, or to compare values within the same data set. The most important concept is the  $z$  score. We will also discuss percentiles and quartiles, as well as a new statistical graph called the boxplot.

### Definition

A  $z$  score (or standardized value) is the number of standard deviations that a given value  $x$  is above or below the mean. The  $z$  score is calculated by using one of the following:

Sample	Population
$z = \frac{x - \bar{x}}{s}$	$z = \frac{x - \mu}{\sigma}$

### Example

Compare those two data values by finding  $z$  score.

Men heights

70.8	66.2	71.7	68.7	67.6	69.2	66.5	67.2	68.3	65.6	63.0	68.3	73.1	67.6
68.0	71.0	61.3	76.2	66.3	69.7	65.4	70.0	62.9	68.5	68.3	69.4	69.2	68.0
71.9	66.1	72.4	73.0	68.0	68.7	70.3	63.7	71.1	65.6	68.3	66.3		

Men weights

169.1	144.2	179.3	175.8	152.6	166.8	135.0	201.5	175.2	139.0	156.3	186.6	191.1
151.3	209.4	237.1	176.7	220.6	166.1	137.4	164.2	162.4	151.8	144.1	204.6	193.8
172.9	161.9	174.8	169.8	213.3	198.0	173.3	214.5	137.1	119.5	189.1	164.7	170.1
151.0												

### Solution

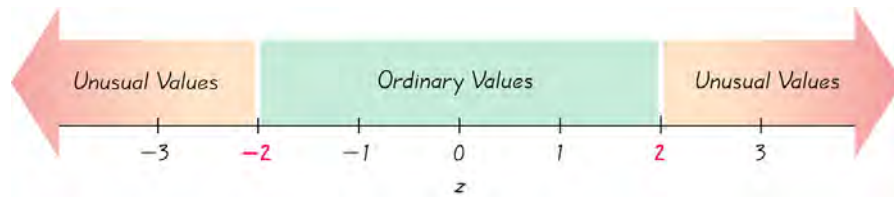
The heights:  $\bar{x} = 68.34$  in. and  $s = 3.02$  in.

$$z = \frac{x - \bar{x}}{s} = \frac{76.2 - 68.34}{3.02} = 2.60$$

The weights:  $\bar{x} = 172.55$  lb. and  $s = 26.33$  lb.

$$z = \frac{x - \bar{x}}{s} = \frac{237.1 - 172.55}{26.33} = 2.45$$

## Z Scores, Unusual Values, and Outliers



Whenever a value is less than the mean, its corresponding  $z$  score is negative

**Ordinary values:**  $-2 \leq z \text{ score} \leq 2$

**Unusual Values:**  $z \text{ score} < -2$  or  $z \text{ score} > 2$

## Percentiles

### Definition

Percentiles are measures of location. There are 99 percentiles denoted  $P_1, P_2, \dots, P_{99}$ , which divide a set of data into 100 groups with about 1% of the values in each group.

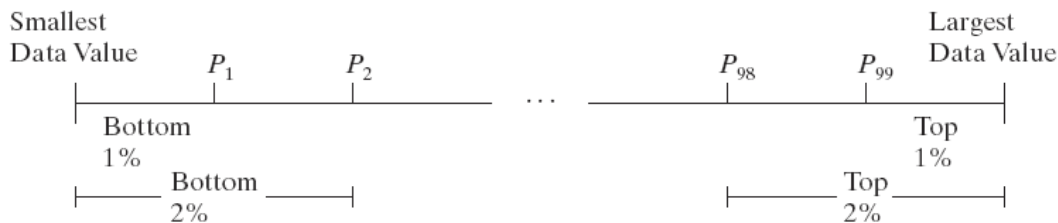
The  $k$ th percentile, denoted,  $P_k$ , of a set of data is a value such that  $k$  percent of the observations are less than or equal to the value.

### Finding the Percentile of a Data Value

The process of finding the percentile that corresponds to a particular data value  $x$  is given by the following:

$$\text{percentile of value } x = \frac{\text{number of values less than } x}{\text{total number of values}} \cdot 100$$

**(Round to the nearest whole number)**



### Example

The table below lists the 35 sorted budget amounts (in millions of dollars) from the simple random sample of movies. Find the percentile for the value of \$29 million

4.5	5	6.5	7	20	20	29	30	35	40	40	41
50	52	60	65	68	68	70	70	70	72	74	75
80	100	113	116	120	125	132	150	160	200	225	

### Solution

From the table, there are 6 budget amounts less than 29, so

$$\text{percentile of } 29 = \frac{6}{35} \cdot 100 \approx 17$$

- ✓ The budget amount of \$29 million is the 17<sup>th</sup> percentile. This can be interpreted loosely as: The budget amount of \$29 million separates the lowest 17% of the budget amounts from the highest 83%.

### Notation

$$L = \frac{k}{100} \cdot n$$

- $n$  total number of values in the data set  
 $k$  percentile being used  
 $L$  locator that gives the position of a value  
 $P_k$   $k$ th percentile

### Example

The table below lists the 35 sorted budget amounts (in millions of dollars) from the simple random sample of movies. Find the value of the 90<sup>th</sup> percentile,  $P_{90}$

4.5	5	6.5	7	20	20	29	30	35	40	40	41
50	52	60	65	68	68	70	70	70	72	74	75
80	100	113	116	120	125	132	150	160	200	225	

### Solution

$k = 90$  and  $n = 35$  because there are 35 data values.

$$\begin{aligned} L &= \frac{k}{100} \cdot n \\ &= \frac{90}{100} \cdot 35 \\ &\approx 32 \end{aligned}$$

The 32nd value is 150 that is,  $P_{90} = \$150$  million.

So, about 90% of the movies have budgets below \$150 million and about 10% of the movies have budgets above \$150 million.

### Example

The list of setting speed limits are recorded (in mi/h.) and listed below

68	68	72	73	65	74	73	72	68	65	65	73	66	71
68	74	66	71	65	73	59	75	70	56	66	75	68	75
62	72	60	73	61	75	58	74	60	73	58	74		

That section has a posted speed limit of 65 mi/h. Traffic engineers often establish speed limits by using the “85<sup>th</sup> percentile rule” whereby the speed limit is set so that 85% of drivers are at or below the speed limit.

- Find the 85<sup>th</sup> percentile of the listed speeds.
- Given that speed limits are usually rounded to a multiple of 5, what speed limit is suggested by these data? Explain your choice.
- Does the existing speed limit conform to the 85<sup>th</sup> percentile rule?

### Solution

Sorting the data

56 58 58 59 60 60 61 62 65 65 65 65 66 66 66 68 68 68 68 68  
70 71 71 72 72 72 73 73 73 73 73 73 74 74 74 74 75 75 75 75

- a)  $n = 40$  because there are 40 sample. To find the 85<sup>th</sup> percentile, then  $k = 85$ .

$$L = \frac{k}{100} \cdot n = \frac{85}{100} \cdot 40 = 34$$

That indicated that the 85<sup>th</sup> percentile is 34<sup>th</sup> speeds, the 34<sup>th</sup> speed is 74 mi/h.

- b) A speed of 75 mi/h is the multiple of 5 closest to the 85<sup>th</sup> percentile, but it is probably safer to round down, so that a speed of 70 mi/h is the closest multiple of 5 below the 85<sup>th</sup> percentile.
- c) The existing speed limit of 65 mi/h is below the speed limit determined by the 85<sup>th</sup> percentile rule, so the existing speed limit does not conform to the 85<sup>th</sup> percentile rule.

### Example

Find  $P_{75}$  for the set of test scores below

51	54	64	68	72	74	76	82	89	94	99
----	----	----	----	----	----	----	----	----	----	----

### Solution

Data already sorted.

$$k = \frac{L}{100} \cdot n = \frac{75}{100} \cdot 11 = 8.25$$

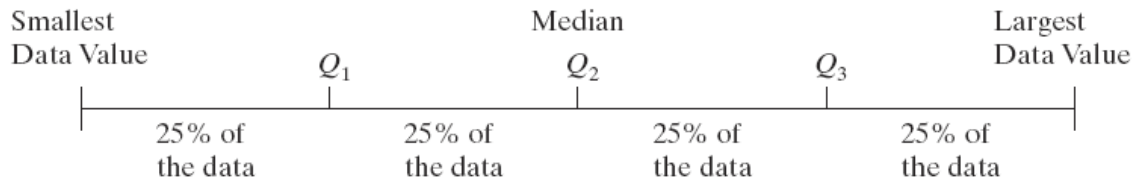
- we find the 8<sup>th</sup> number in the list = 82
- $.25(9^{\text{th}} \# - 8^{\text{th}} \#) = .25(89 - 82) = 1.75$
- $P_{75} = 82 + 1.75 = 83.75$

## Quartiles

### Definition

Quartile are measures of location, denoted  $Q_1$ ,  $Q_2$ , and  $Q_3$ , which divide a set of data into four groups with about 25% of the values in each group.

- $Q_1$  (First Quartile) separates the bottom 25% of sorted values from the top 75%.
- $Q_2$  (Second Quartile) same as the median; separates the bottom 50% of sorted values from the top 50%.
- $Q_3$  (Third Quartile) separates the bottom 75% of sorted values from the top 25%.



### Example

Find the value of the first quartile  $Q_1$ .

4.5	5	6.5	7	20	20	29	30	35	40	40	41
50	52	60	65	68	68	70	70	70	72	74	75
80	100	113	116	120	125	132	150	160	200	225	

### Solution

Finding  $Q_1$  is the same as finding  $P_{25}$

$$L = \frac{k}{100} \cdot n = \frac{25}{100} \cdot 35 = 8.75 \approx 9.$$

Therefore, the first quartile is given by  $Q_1 = \$35$  million.

- ✓ Interquartile range (or **IRQ**) =  $Q_3 - Q_1$
- ✓ Semi-interquartile range =  $\frac{Q_3 - Q_1}{2}$
- ✓ Midquartile =  $\frac{Q_3 + Q_1}{2}$
- ✓ 10–90 percentile range =  $P_{90} - P_{10}$

### Definition

For a set of data, the **5-number summary** consists of the minimum value; the first quartile  $Q_1$ ; the median (or second quartile  $Q_2$ ); the third quartile,  $Q_3$ ; and the maximum value.

A **boxplot** (or **box-and-whisker-diagram**) is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile,  $Q_1$ ; the median; and the third quartile,  $Q_3$ .

## Example

Use the movie budget amount to find the 5-number summary.

4.5	5	6.5	7	20	20	29	30	35	40	40	41
50	52	60	65	68	68	70	70	70	72	74	75
80	100	113	116	120	125	132	150	160	200	225	

## Solution

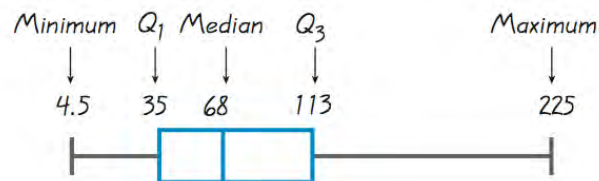
From the table:

The minimum is \$4.5 million and the maximum is \$225 million.

$$L = \frac{25}{100} \cdot 35 \approx 9 \Rightarrow P_{25} = 35 \rightarrow Q_1 = \$35 \text{ million}$$

$$L = \frac{50}{100} \cdot 35 \approx 18 \Rightarrow P_{50} = 68 \rightarrow Q_2 = \$68 \text{ million.}$$

$$L = \frac{75}{100} \cdot 35 \approx 27 \Rightarrow P_{75} = 113 \rightarrow Q_3 = \$113 \text{ million.}$$



## Procedure for Constructing a Boxplot

1. Find the 5-number summary consisting of the minimum value,  $Q_1$ , the median,  $Q_3$ , and the maximum value.
2. Construct a scale with values that include the minimum and maximum data values.
3. Construct a box (rectangle) extending from  $Q_1$  to  $Q_3$ , and draw a line in the box at the median value.
4. Draw lines extending outward from the box to the minimum and maximum data values.

## Outliers and Modified Boxplots

### Definition

An **outlier** is a value that lies very far away from the vast majority of the other values in a data set.

For purposes of constructing *modified boxplots*, we can consider outliers to be data values meeting specific criteria.

In modified boxplots, a data value is an outlier if it is . . .

above  $Q_3$  by an amount greater than  $1.5 \times \text{IQR}$

or

below  $Q_1$  by an amount greater than  $1.5 \times \text{IQR}$

### Example

The pulse rates of females listed below

76	72	88	60	72	68	80	64	68	68	80	76	68	72	96
72	68	72	64	80	64	80	76	76	76	80	104	88	60	76
72	72	88	80	60	72	88	88	124	64					

Use the data to construct a modified boxplot.

### Solution

60	60	60	64	64	64	64	68	68	68	68	68	72	72	72	72
72	72	72	72	76	76	76	76	76	76	80	80	80	80	80	80
88	88	88	88	88	96	104	124								

From the table: The minimum is 60 and the maximum is 124.

$$L = \frac{25}{100} \cdot 40 \approx 10 \Rightarrow P_{25} = 68 \rightarrow Q_1 = 68$$

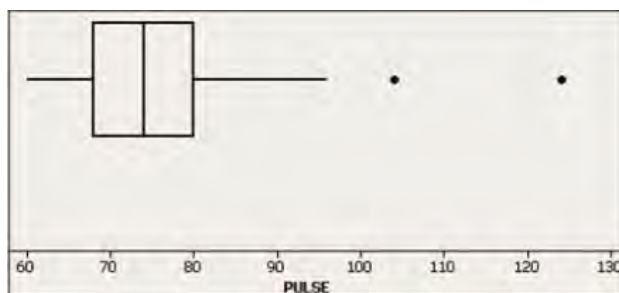
$$L = \frac{75}{100} \cdot 40 \approx 30 \Rightarrow P_{75} = 80 \rightarrow Q_3 = 80$$

Interquartile range (or ***IRQ***) =  $Q_3 - Q_1 = 80 - 68 = 12$

For pulse rates above the third quartile of 80 by an amount that is greater than

$1.5 \times IRQ = 1.5 \times 12 = 18$  so high outliers are greater than 98.

The pulse rates of 104 and 124 satisfy this condition, so those two values are outliers.



Store the list values in L1.

Press [2<sup>nd</sup>] (Y=) [STAT PLOT] [1]

Press [ENTER] to turn on the stat plot.

Scroll down to Type, then right 4 times first pictures in 2<sup>nd</sup> row.

Be sure that Xlist is L1

Allow Freq: 1

Press [GRAPH], Press [ZOOM] and select 9, and press [ENTER]

## Exercises    **Section 1.10 – Measures of Position, Outliers, and Boxplots**

1. When Reese Witherspoon won an Oscar as Best Actress for the movie *Walk the Line*, her age was converted to a z-score of  $-0.61$  when included among the ages of all other Oscar-winning Best Actress at the time of this writing. Was her age above the mean or below the mean? How many standard deviations away from the mean is her age?
2. Hoffman was 38 years of age when he won a Best Actor Oscar for his role in *Capote*. The Oscar-winning Best Actors have a mean age of 43.8 years and a standard deviation of 8.9 years.
  - a) What is the difference between Hoffman's age and the mean age?
  - b) How many standard deviations is that (the difference found in part (a))?
  - c) Convert Hoffman's age to a z-score.
  - d) If we consider "usual" ages to be those that convert to z-scores between  $-2$  and  $2$ , is Hoffman's age usual or unusual?
3. Eruptions of the Old Faithful geyser have duration times with a mean of 245.0 sec and a standard deviation of 36.4 sec. One eruption had a duration time of 110 sec.
  - a) What is the difference between a duration time of 110 sec and the mean?
  - b) How many standard deviations is that (the difference found in part (a))?
  - c) Convert duration time of 110 sec to a z-score.
  - d) If we consider "usual" ages to be those that convert to z-scores between  $-2$  and  $2$ , is a duration time of 110 sec usual or unusual?
4. Human body temperatures have a mean of  $98.20^{\circ}\text{F}$  and a standard deviation of  $0.62^{\circ}\text{F}$ . Convert each given temperature to a z-score and determine whether it is usual and unusual.
  - a)  $101.00^{\circ}\text{F}$
  - b)  $96.90^{\circ}\text{F}$
  - c)  $96.98^{\circ}\text{F}$
5. Scores on SAT test have a mean of 1518 and a standard deviation of 325. Scores on the ACT test have a mean of 21.1 and standard deviation of 4.8. Which is relatively better: a score of 1840 on the SAT test or a score of 26.0 on the ACT test? Why?
6. Scores on SAT test have a mean of 1518 and a standard deviation of 325. Scores on the ACT test have a mean of 21.1 and standard deviation of 4.8. Which is relatively better: a score of 1190 on the SAT test or a score of 16.0 on the ACT test? Why?
7. Use the given sorted values, which are the numbers of points scored in the Super Bowl for a recent period of 24 years. Find the percentile corresponding to the given number of points  
36 37 37 39 39 41 43 44 44 47 50 53 54 55 56 56 57 59 61 61 65 69 69 75
  - a) 47
  - b) 65
  - c) 54
  - d) 41



8. For the given data, find the indicated percentile or quartile

36 37 37 39 39 41 43 44 44 47 50 53 54 55 56 56 57 59 61 61 65 69 69 75

a)  $P_{20}$

c)  $P_{50}$

e)  $P_{25}$

g)  $Q_1$

b)  $P_{80}$

d)  $P_{75}$

f)  $P_{95}$

h)  $Q_3$

9. The number of hours of television watched per day by a sample of 28 people

2 4 1 5 7 2 5 4 4 2 3 6 4 3 5 2 0 3 5 9 4 5 2 1 3 6 7 2

- Find the data set's first, second, and third quartiles.
- Draw a box-and-whisker plot that represents the data set.
- About 75% of the people watched no more than how many hours of television per day?
- What percent of the people watched more than 4 hours of television per day?
- If you randomly selected one person from the sample, what is the likelihood that the person watched less than 2 hours of television per day? Write your answer as a percent.

10. The hourly earnings (in dollars) of a sample of 25 railroad equipment manufacturers

15.6 18.75 14.6 15.8 14.35 13.9 17.5 17.55 13. 14.2 19.05 15.35 15.2  
19.45 15.95 16.5 16.3 15.25 15.05 19.1 15.2 16.22 17.75 18.4 15.25

- Find the data set's first, second, and third quartiles.
- Draw a box-and-whisker plot that represents the data set.
- About 75% of the manufacturers made less than \$15.80 per hour?
- What percent of the manufacturers made more than \$15.80 per hour?
- If you randomly selected one manufacturer from the sample, what is the likelihood that the manufacturer made less than \$15.80 per hour? Write your answer as a percent.

11. A certain brand of automobile tire has a mean life span of 35,000 miles, with a standard deviation of 2250 miles. (Assume the life spans of the tires have a bell-shaped distribution)

- The life spans of three randomly selected tires are 34,000 miles, 37,000 miles, and 30,000 miles. Find the  $z$ -score that corresponds to each life span. According to the  $z$ -scores, would the life spans of any of these tires be considered unusual?
- The life spans of three randomly selected tires are 30,500 miles, 37,250 miles, and 35,000 miles. Using the Empirical Rule, find the percentile that corresponds to each life span.

12. The life spans of species of fruit fly have a bell shaped distribution, with mean of 33 days and a standard deviation of 4 days.

- The life spans of three randomly selected fruit flies are 34 days, 30 days, and 42 days. Find the  $z$ -score that corresponds to each life span and determine if any of these life spans are unusual.
- The life spans of three randomly selected fruit flies are 29 days, 41 days, and 25 days. Using the Empirical Rule, find the percentile that corresponds to each life span.

13. Find the  $Q_1$  and  $Q_3$  for the given data: 49 52 52 52 74 67 55 55

14. Find the  $Q_1$  and  $Q_3$  for the given weights (in pounds) of 30 newborn babies listed below:

5.5 5.7 5.8 6.0 6.1 6.1 6.3 6.4 6.5 6.6  
6.7 6.7 6.7 6.9 7.0 7.0 7.0 7.1 7.2 7.2  
7.4 7.5 7.7 7.7 7.8 8.0 8.1 8.1 8.3 8.7

15. Find the percentile for the data value:

113 125 117 111 119 121 111 109 116 113 117 127 109 113 115 110  
Data value: 119

16. The test scores of 40 students are listed below:

30 35 43 44 47 48 54 55 56 57 59 62 63 65 66 68 69 69 71 72  
72 73 74 76 77 77 78 79 80 81 81 82 83 85 89 92 93 94 97 98

Find  $P_{56}$