STATISTICS
INFORMED DECISIONS USING DATA
Fifth Edition

Tuesday Feb 18

Chapter 1
Data Collection

Pearson

1.1 Introduction to the Practice of Statistics

Learning Objectives

1. Define statistics and statistical thinking
2. Explain the process of statistics
3. Distinguish between qualitative and quantitative variables
4. Distinguish between discrete and continuous variables
5. Determine the level of measurement of a variable

1.1 Introduction to the Practice of Statistics

1.1.1 Define Statistics and Statistical Thinking

Statistics is …

1.1 Introduction to the Practice of Statistics

1.1.1 Define Statistics and Statistical Thinking

Statistics is the science of collecting, organizing, summarizing, and analyzing information to draw conclusions or answer questions. In addition, statistics is about providing a measure of confidence in any conclusions.

1.1 Introduction to the Practice of Statistics

1.1.1 Define Statistics and Statistical Thinking

Statistics is the science of collecting, organizing, summarizing, and analyzing information to draw conclusions or answer questions. In addition, statistics is about providing a measure of confidence in any conclusions.

The information referred to in the definition is **data**. **Data** are a "fact or proposition used to draw a conclusion or make a decision." Data describe characteristics of an individual.

1.1 Introduction to the Practice of Statistics

1.1.1 Define Statistics and Statistical Thinking

Statistics is the science of collecting, organizing, summarizing, and analyzing information to draw conclusions or answer questions. In addition, statistics is about providing a measure of confidence in any conclusions.

The information referred to in the definition is **data**. **Data** 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.

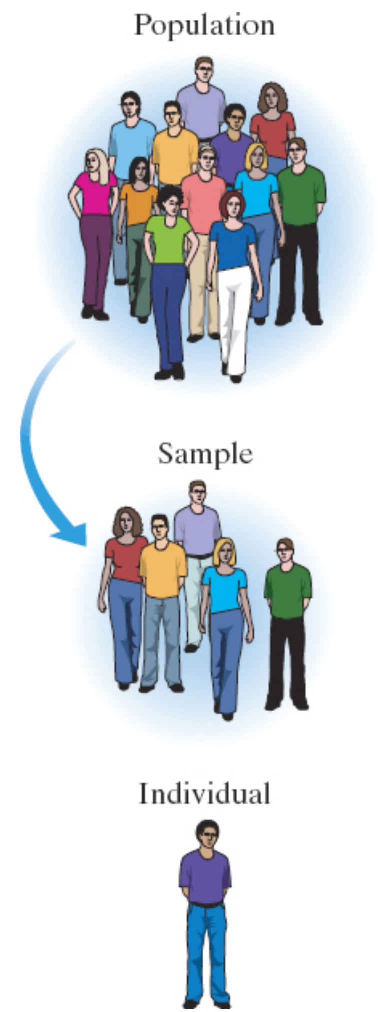
In fact, data vary when measured on ourselves as well. Do you sleep the same number of hours every night? No! Do you consume the same number of calories every day? No!

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

1.1 Introduction to the Practice of Statistics

1.1.2 Explain the Process of Statistics (1 of 7)

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 population vs sample
(all) (some)

1.1 Introduction to the Practice of Statistics

1.1.2 Explain the Process of Statistics (2 of 7)

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.

Ex: average height of all GCC students

A **parameter** is a numerical characteristic of a population. (PP)

A **statistic** is a numerical characteristic of a sample. (SS)

Ex average height of GCC students in Jorge's 136 class

1.1 Introduction to the Practice of Statistics
1.1.2 Explain the Process of Statistics (3 of 7)

EXAMPLE Parameter versus Statistic

Suppose the percentage of all students on your campus who have a job is 84.9%. This value represents a parameter because it is a numerical characteristic of a population.

Suppose a sample of 250 students is obtained, and from this sample we find that 86.4% have a job. This value represents a statistic because it is a numerical characteristic based on a sample.

1.1 Introduction to the Practice of Statistics

1.1.2 Explain the Process of Statistics

The Process of Statistics

1. **Identify the research objective**. A researcher must determine the question(s) he or she wants answered. The question(s) must clearly identify the population that is to be studied.

2. **Collect the data needed to answer the question(s) posed in (1).** Conducting research on an entire population is often difficult and expensive, so we typically look at a sample. This step is vital to the statistical process, because if the data are not collected correctly, the conclusions drawn are meaningless. Do not overlook the importance of appropriate data collection. We discuss this step in detail in Sections 1.2 through 1.6.

3. **Describe the data.** Descriptive statistics allow the researcher to obtain an overview of the data and can help determine the type of statistical methods the researcher should use. We discuss this step in detail in Chapters 2 through 4.

4. **Perform inference.** Apply the appropriate techniques to extend the results obtained from the sample to the population and report a level of reliability of the results. We discuss techniques for measuring reliability in Chapters 5 through 8 and inferential techniques in Chapters 9 through 15.

1.1 Introduction to the Practice of Statistics
1.1.2 Explain the Process of Statistics (5 of 7)

EXAMPLE Illustrating the Process of Statistics

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.

Source: *The Effects of a Group Batterer Treatment Program: A Randomized Experiment in Brooklyn by Bruce G. Taylor, et. al. Justice Quarterly, Vol. 18, No. 1, March 2001.*

1.1 Introduction to the Practice of Statistics

1.1.2 Explain the Process of Statistics (6 of 7)

Step 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.

Step 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.

1.1 Introduction to the Practice of Statistics
1.1.2 Explain the Process of Statistics (7 of 7)

Step 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.

Step 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.

1.1 Introduction to the Practice of Statistics

1.1.3 Distinguish between Qualitative and Quantitative Variables (1 of 3)

Wed Feb 19

Ex height of a single OCC student

Variables are the characteristics of the individuals within the individuals of a 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.

1.1 Introduction to the Practice of Statistics

1.1.3 Distinguish between Qualitative and Quantitative Variables (2 of 3)

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

1.1 Introduction to the Practice of Statistics

1.1.3 Distinguish between Qualitative and Quantitative Variables (2 of 3)

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

(QL)

Qualitative variables provide a description or category of individuals. The values of a qualitative variables cannot be added or subtracted and provide meaningful results.

vs

(QN)

Quantitative variables provide **numerical** measures of individuals. The values of a quantitative variable can be added or subtracted and provide meaningful results.

EXAMPLE Distinguishing between Qualitative and Quantitative Variables

Researcher Elisabeth Kvaavik and others studied factors that affect the eating habits of adults in their mid-thirties. (Source: Kvaavik E, et. al. Psychological explanatorys of eating habits among adults in their mid-30's (2005) International Journal of Behavioral Nutrition and Physical Activity (2)9.)

Classify each of the following variables considered in the study as qualitative (QL) or quantitative (QN).

a. Nationality QL
b. Number of children QN
c. Household income in the previous year QN
d. Level of education QL
e. Daily intake of whole grains (measured in grams per day) QN

1.1 Introduction to the Practice of Statistics

1.1.3 Distinguish between Qualitative and Quantitative Variables (3 of 3)

EXAMPLE Distinguishing between Qualitative and Quantitative Variables

Researcher Elisabeth Kvaavik and others studied factors that affect the eating habits of adults in their mid-thirties. (Source: Kvaavik E, et. al. Psychological explanatorys of eating habits among adults in their mid-30's (2005) International Journal of Behavioral Nutrition and Physical Activity (2)9.)

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

a. Nationality **Qualitative**

b. Number of children **Quantitative**

c. Household income in the previous year **Quantitative**

d. Level of education **Qualitative**

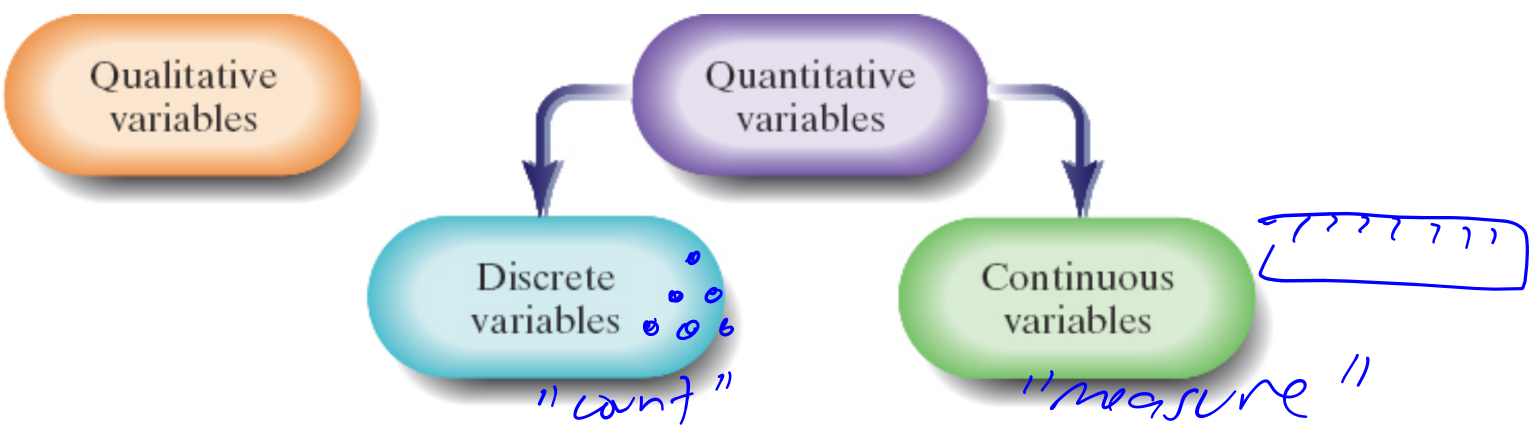
e. Daily intake of whole grains (measured in grams per day) **Quantitative**

1.1 Introduction to the Practice of Statistics

1.1.4 Distinguish between Discrete and Continuous Variables (1 of 3)

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. A discrete variable cannot take on every possible value between any two possible values.

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.



Pearson

1.1 Introduction to the Practice of Statistics

1.1.4 Distinguish between Discrete and Continuous Variables (2 of 3)

EXAMPLE Distinguishing between Discrete and Continuous Variables

Researcher Elisabeth Kvaavik and others studied factors that affect the eating habits of adults in their mid-thirties. (Source: Kvaavik E, et. al. Psychological explanatorys of eating habits among adults in their mid-30's (2005) International Journal of Behavioral Nutrition and Physical Activity (2)9.)

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

a. Number of children discrete

b. Household income in the previous year continual

c. Daily intake of whole grains (measured in grams per day) continuous

Pearson

EXAMPLE Distinguishing between Discrete and Continuous Variables

Researcher Elisabeth Kvaavik and others studied factors that affect the eating habits of adults in their mid-thirties. (Source: Kvaavik E, et. al. Psychological explanatorys of eating habits among adults in their mid-30's (2005) International Journal of Behavioral Nutrition and Physical Activity (2)9.)

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

a. Number of children **Discrete**

b. Household income in the previous year **Continuous**

c. Daily intake of whole grains (measured in grams per day) **Continuous**

1.1 Introduction to the Practice of Statistics

1.1.4 Distinguish between Discrete and Continuous Variables (3 of 3)

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.

1.1 Introduction to the Practice of Statistics

1.1.5 Determine the Level of Measurement of a Variable (1 of 3)

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.

Quick Example: Names of Colleges visited (but not ranked).

Levels of Measurement

- nominal
- ordinal
- interval
- ratio

1.1 Introduction to the Practice of Statistics

1.1.5 Determine the Level of Measurement of a Variable (1 of 3)

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.

Quick Example: 1rst prize winner, 2^{nd} prize winner, etc

1.1 Introduction to the Practice of Statistics

1.1.5 Determine the Level of Measurement of a Variable (2 of 3)

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.

Quick Example: Temperatures! 0 degrees F is not the absence of degrees (it can get way colder than 0 degrees F)

1.1 Introduction to the Practice of Statistics

1.1.5 Determine the Level of Measurement of a Variable (2 of 3)

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.

Quick Example: You name it—this one is the easiest!

1.1 Introduction to the Practice of Statistics

1.1.5 Determine the Level of Measurement of a Variable (3 of 3)

EXAMPLE Determining the Level of Measurement of a Variable

A study was conducted to assess school eating patterns in high schools in the United States. The study analyzed the impact of vending machines and school policies on student food consumption. A total of 1088 students in 20 schools were surveyed. (Source: Neumark-Sztainer D, French SA, Hannan PJ, Story M and Fulkerson JA (2005) School lunch and snacking patterns among high school students: associations with school food environment and policies. International Journal of Behavioral Nutrition and Physical Activity 2005, (2)14.)

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

a. Number of snack and soft drink vending machines in the school ratio — true zero, can have zero machines

b. Whether or not the school has a closed campus policy during lunch nominal

c. Class rank (Freshman, Sophomore, Junior, Senior) ordinal

d. Number of days per week a student eats school lunch ratio — true zero, can have 0 days of lunch

Pearson

1.1 Introduction to the Practice of Statistics

1.1.5 Determine the Level of Measurement of a Variable (3 of 3)

EXAMPLE Determining the Level of Measurement of a Variable

A study was conducted to assess school eating patterns in high schools in the United States. The study analyzed the impact of vending machines and school policies on student food consumption. A total of 1088 students in 20 schools were surveyed. (Source: Neumark-Sztainer D, French SA, Hannan PJ, Story M and Fulkerson JA (2005) School lunch and snacking patterns among high school students: associations with school food environment and policies. International Journal of Behavioral Nutrition and Physical Activity 2005, (2)14.)

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

a. Number of snack and soft drink vending machines in the school **Ratio**

b. Whether or not the school has a closed campus policy during lunch **Nominal**

c. Class rank (Freshman, Sophomore, Junior, Senior) **Ordinal**

d. Number of days per week a student eats school lunch **Ratio**

1.2 Observational Studies Versus Designed Experiments
Learning Objectives

1. Distinguish between an observational study and an experiment
2. Explain the various types of observational studies

Compare next two studies…

1.2 Observational Studies Versus Designed Experiments

1.2.1 Distinguish between an observational study and an experiment (1 of 10)

EXAMPLE Cellular Phones and Brain Tumors, study 1

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.

1.2 Observational Studies Versus Designed Experiments

1.2.1 Distinguish between an observational study and an experiment (2 of 10)

EXAMPLE Cellular Phones and Brain Tumors, study 1

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)

This is observational study

researchers don't try to influence people in study

EXAMPLE Cellular Phones and Brain Tumors, study 2

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.

EXAMPLE Cellular Phones and Brain Tumors, study 2

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.)

This is a designed study

researchers do try to influence rats in study (in Groups 1 & 2)

Pearson

1.2 Observational Studies Versus Designed Experiments

1.2.1 Distinguish between an observational study and an experiment

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 (dependent variable)**.

The level of cell phone usage is the **explanatory variable (independent variable)**.

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

1.2 Observational Studies Versus Designed Experiments

1.2.1 Distinguish between an observational study and an experiment (6 of 10)

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.

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**.

1.2 Observational Studies Versus Designed Experiments
1.2.1 Distinguish between an observational study and an experiment (7 of 10)

EXAMPLE Observational Study or Designed Experiment? Do Flu shots Benefit Seniors?

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)

1.2 Observational Studies Versus Designed Experiments
1.2.1 Distinguish between an observational study and an experiment (8 of 10)

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**.

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 affects the value of the response variable in the study. In addition, lurking variables are typically related to any explanatory variables considered in the study.

1.2 Observational Studies Versus Designed Experiments

1.2.1 Distinguish between an observational study and an experiment (9 of 10)

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.

1.2 Observational Studies Versus Designed Experiments

1.2.1 Distinguish between an observational study and an experiment (10 of 10)

Observational studies do not allow a researcher to claim causation, only association.

A **confounding variable** is an explanatory variable that was considered in a study whose effect cannot be distinguished from a second explanatory variable in the study.

Monday 24

1.2 Observational Studies Versus Designed Experiments

1.2.2 Explain the Various Types of Observational Studies (1 of 6)

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

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. (ex: did your ancestors come to US in last 50 yrs).

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 Observational Study or Designed Experiment?

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: case-control

EXAMPLE Observational Study or Designed Experiment?

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

1.2.2 Explain the Various Types of Observational Studies (3 of 6)

EXAMPLE Observational Study or Designed Experiment?

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.)

Experiment : cross-sectional

EXAMPLE Observational Study or Designed Experiment?

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, cross-sectional

1.2 Observational Studies Versus Designed Experiments
1.2.2 Explain the Various Types of Observational Studies (4 of 6)

EXAMPLE Observational Study or 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.)

EXAMPLE Observational Study or 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

EXAMPLE Observational Study or Designed Experiment?

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)

1.2 Observational Studies Versus Designed Experiments

1.2.2 Explain the Various Types of Observational Studies (5 of 6)

EXAMPLE Observational Study or Designed Experiment?

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

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

Very expensive and hard to achieve perfection.

Government do these only every 10 years.

1.3 Simple Random Sampling

Learning Objectives

1. Obtain a simple random sample

Super Important!

1.3 Simple Random Sampling

1.3.1 Obtain a simple random sample (1 of 8)

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.

true random sample:
every individual has equal chance being selected.

1.3 Simple Random Sampling

1.3.1 Obtain a simple random sample (2 of 8)

A sample of size *n* from a population of size *N* is obtained through **simple random sampling (SRS)** if every possible sample of size *n* has an equally likely chance of occurring.

The sample is then called a **simple random sample (SRS).**

SRS

- sample of size n
- each person equal chance of being selected.

1.3 Simple Random Sampling

1.3.1 Obtain a simple random sample (3 of 8)

EXAMPLE Illustrating Simple Random Sampling

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).

- Bob, Pat
- Bob, Mike
- Bob, Jan
- Bob, Maria

Pat, Mike
Pat, Jan
Pat, Maria

Mike Jan
Mike Maria

Jan Maria

Note there are 10 SRS

1.3 Simple Random Sampling

1.3.1 Obtain a simple random sample (3 of 8)

EXAMPLE Illustrating Simple Random Sampling

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).

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

1.3 Simple Random Sampling

1.3.1 Obtain a simple random sample (5 of 8)

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 to *N*.

2) Use a random number table, graphing calculator, or statistical software to randomly generate *n* numbers where *n* is the desired sample size.

notation

N – population size

n – sample size

1.3 Simple Random Sampling

1.3.1 Obtain a simple random sample (6 of 8)

EXAMPLE Obtaining a Simple Random Sample

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~~.

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

list

1.3 Simple Random Sampling

1.3.1 Obtain a simple random sample (7 of 8)

EXAMPLE Obtaining a Simple Random Sample

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.

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.

#1: 206 #2: 215 #3: 5 #4: 179 #5: 381

1.3 Simple Random Sampling

1.3.1 Obtain a simple random sample (8 of 8)

EXAMPLE Obtaining a Simple Random Sample

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.

Step 3 Match the generated random numbers to the corresponding Representatives.

1.4 Other Effective Sampling Methods

Learning Objectives

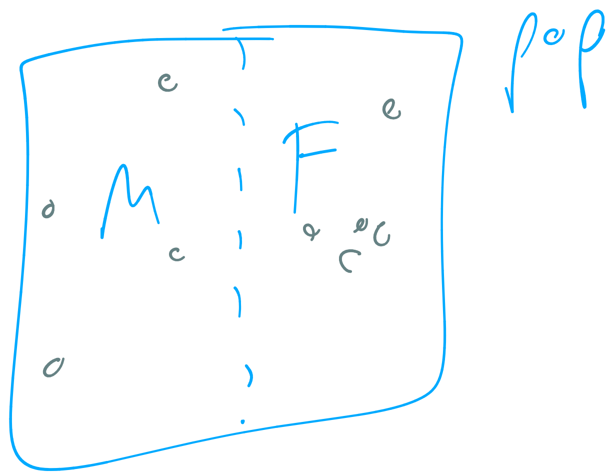
1. Obtain a stratified sample
2. Obtain a systematic sample
3. Obtain a cluster sample

1.4 Other Effective Sampling Methods

1.4.1 Obtain a stratified Sample (1 of 2)

A **stratified sample** is obtained by separating the population into 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.

Quick Example: Split into Males and Females and select SRS from each.



1.4 Other Effective Sampling Methods

1.4.1 Obtain a stratified Sample (2 of 2)

EXAMPLE Obtaining a Stratified Sample

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.

Gp 1
Use RNG
pick
4 R

Gp 2
Use
~~RNG~~
pick
4 D

Gp 3
pick one Ind.

1.4 Other Effective Sampling Methods

1.4.1 Obtain a stratified Sample (2 of 2)

EXAMPLE Obtaining a Stratified Sample

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.

1.4 Other Effective Sampling Methods

1.4.2 Obtain a Systematic Sample (1 of 3)

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

Quick Example: Select at radom a number between 1 and 10. Then select the 10th, 20th, 30th, etc (k=10)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 ......

k

k=3

Pearson

1.4 Other Effective Sampling Methods
1.4.2 Obtain a Systematic Sample (2 of 3)

EXAMPLE Obtaining a Systematic Sample

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.

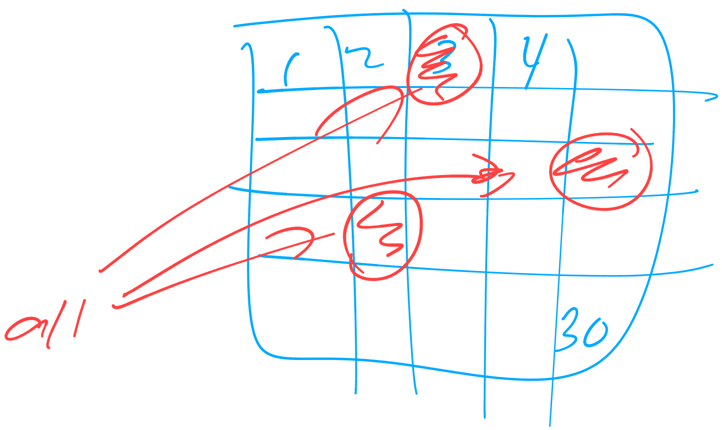
· randomly select k = 4,

1.4 Other Effective Sampling Methods

1.4.3 Obtain a Cluster Sample (1 of 7)

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

Quick Example: Split a district into 30 zones (geographically) and select randomly 5 zones. Then everyone from these 5 zones is part of a cluster sample.



Randomly sample groups

(here: 3 groups)

key: select all individuals in each group selected

Pearson

1.4 Other Effective Sampling Methods

1.4.3 Obtain a Cluster Sample (2 of 7)

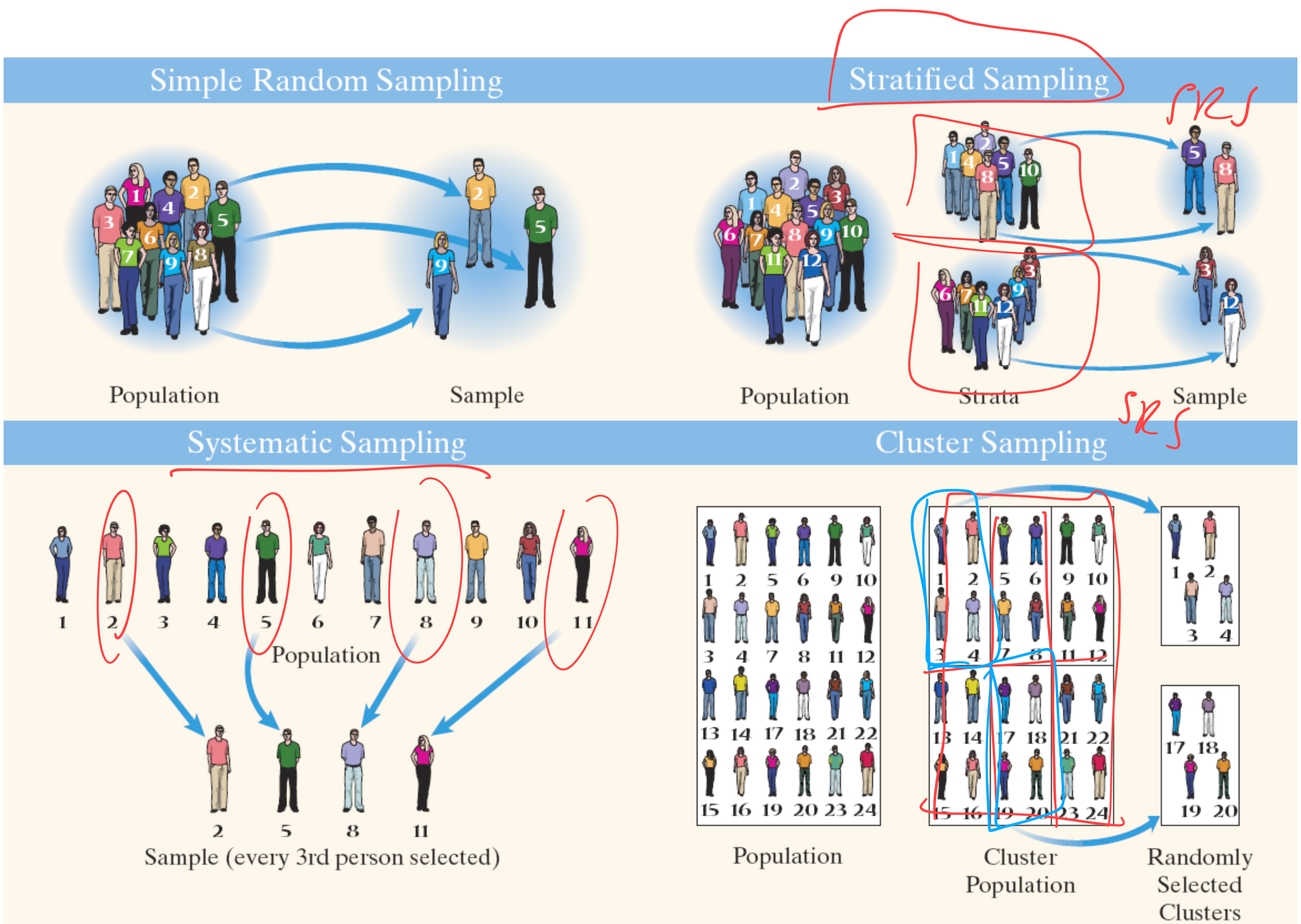
EXAMPLE Obtaining a Cluster Sample

A school administrator wants to obtain a sample of students in order to conduct a survey.

She randomly selects 10 classes and administers the survey to all the students in the class.

1.4 Other Effective Sampling Methods

1.4.3 Obtain a Cluster Sample (3 of 7)



1.4 Other Effective Sampling Methods

1.4.3 Obtain a Cluster Sample (4 of 7)

Stratified and cluster samples are different.

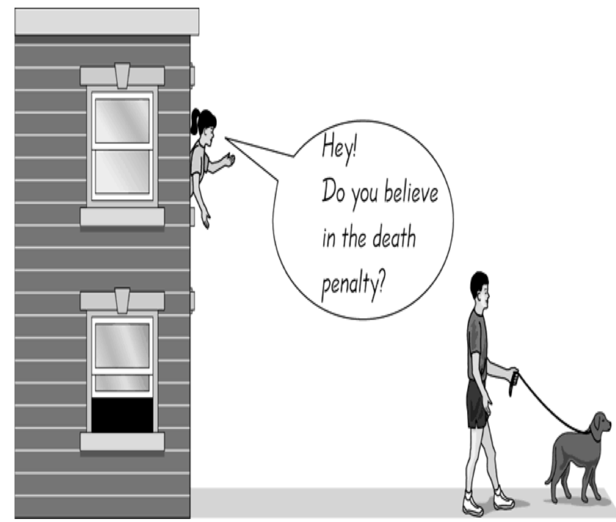
In a stratified sample, we divide the population into two or more homogeneous groups. Then we obtain a simple random sample from each group.

In a cluster sample, we divide the population into groups, obtain a simple random sample of some of the groups, and survey **all** individuals in the selected groups.

1.4 Other Effective Sampling Methods

1.4.3 Obtain a Cluster Sample (5 of 7)

A **convenience sample** is one in which the individuals in the sample are easily obtained.



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

1.5 Bias in Sampling

Learning Objectives

1. Explain the sources of bias in sampling

1.5 Bias in Sampling

1.5.1 Explain the sources of bias in sampling (1 of 6)

If the results of the sample are not representative of the population, then the sample has **bias**.

Three Sources of bias

1. Sampling bias
2. Nonresponse bias
3. Response bias

1.5 Bias in Sampling

1.5.1 Explain the sources of bias in sampling (2 of 6)

Sampling bias means that the technique used to obtain the individuals to be in the sample tends to favor one part of the population over another.

1.5 Bias in Sampling

1.5.1 Explain the sources of bias in sampling (2 of 6)

Sampling bias means that the technique used to obtain the individuals to be in the sample tends to favor one part of the population over another.

Undercoverage results in sampling bias. **Undercoverage** occurs when the proportion of one segment of the population is lower in a sample than it is in the population.

1.5 Bias in Sampling

1.5.1 Explain the sources of bias in sampling (3 of 6)

Nonresponse bias exists when individuals selected to be in the sample who do not respond to the survey have different opinions from those who do.

Nonresponse can be improved through the use of callbacks or rewards/incentives.

1.5 Bias in Sampling

1.5.1 Explain the sources of bias in sampling (4 of 6)

Response bias exists when the answers on a survey do not reflect the true feelings of the respondent.

Types of Response Bias

1. Interviewer error
2. Misrepresented answers
3. Wording of questions
4. Order of questions or words

1.5 Bias in Sampling

1.5.1 Explain the sources of bias in sampling (6 of 6)

Sampling error is an error that results from using a sample to estimate information about a population. This type of error occurs because a sample gives incomplete information about a population.

Tues 2/25

1.6 The Design of Experiments
Learning Objectives

1. Describe the characteristics of an experiment
2. Explain the steps in designing an experiment
3. Explain the completely randomized design
4. Explain the matched-pairs design
5. Explain the randomized block design

1.6 The Design of Experiments

1.6.1 Describe the characteristics of an experiment (1 of 4)

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.

"sugar pill"

1.6 The Design of Experiments

1.6.1 Describe the characteristics of an experiment (2 of 4)

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

→ reduce: placebo effect

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.

1.6 The Design of Experiments

1.6.1 Describe the characteristics of an experiment (3 of 4)

EXAMPLE The Characteristics of an Experiment

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.

→ subjects

(a) Who are the experimental units?

The students in the class

(b) What is the population for which this study applies?

The students in the class

1.6 The Design of Experiments

1.6.1 Describe the characteristics of an experiment (4 of 4)

(c) What are the treatments?

Traditional vs. online instruction

(d) What is the response variable?

Exam score

(e) Why can’t this experiment be conducted with blinding?

Both the students and instructor know which treatment they are receiving

1.6 The Design of Experiments

1.6.2 Explain the Steps in Designing an Experiment (1 of 9)

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

1.6 The Design of Experiments

1.6.3 Explain the Completely Randomized Design (1 of 6)

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

1.6 The Design of Experiments

1.6.4 Explain the Matched-Pairs Design (1 of 3)

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.

- same person: height and age
- twins
- couples: sleep
- geography: CA vs NY

1.6 The Design of Experiments

1.6.4 Explain the Matched-Pairs Design (2 of 3)

EXAMPLE A Matched-Pairs Design

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. The researchers measured the children's ratings of the two types of milk. (Source: Castillo JL, et al (2005) *Children's acceptance of milk with Xylitol or Sorbitol for dental caries prevention*. BMC Oral Health (5)6.)

a) What is the response variable in this experiment? **Rating**

b) Think of some of the factors in the study. Which are controlled? Which factor is manipulated?
Age and gender of the children; Milk with and without Xylitol is the factor that was manipulated

1.6 The Design of Experiments

1.6.4 Explain the Matched-Pairs Design (3 of 3)

c) What are the treatments? How many treatments are there?
Milk with Xylitol and milk without xylitol; 2

d) What type of experimental design is this?
Matched-pairs design

e) Identify the experimental units. **75 Peruvian children**

f) Why would it be a good idea to randomly assign whether the child drinks the milk with Xylitol first or second?
Remove any effect due to order in which milk is drunk.

g) Do you think it would be a good idea to double-blind this experiment? **Yes!**

1.6 The Design of Experiments

1.6.5 Explain the Randomized Block Design (1 of 4)

Grouping together similar (homogeneous) experimental units and then randomly assigning 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.

ie cant split between
explanatory & response.

1.6 The Design of Experiments

1.6.5 Explain the Randomized Block Design (2 of 4)

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.