**Section 1.1 Introduction to the Practice of Statistics**

**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

***Objective 1: Define Statistics and Statistical Thinking***

Objective 1, Page 1

 *Answer the following as you watch the video.*

1. Write the definition of statistics below the science of collect , organize, summarize, and analyze, information to draw conclusions or answer questions. In addition, stats is about providing a MEASURE of confidence in any conclustion.
2. Data describes CHARACTERISTICS of individuals and can be either group or within the individual himself.

**Note:** Data varies. Consider the students in your class. Is everyone the same height? No. Does everyone have the same color hair? No. So, within groups there is variation. Now consider yourself. Do you eat the same amount of food (as measured by calories) each day? No. Do you sleep the same number of hours each day? No. So, even considering individuals there is variation. One goal of statistics is to describe and understand sources of variation.

***Objective 2: Explain the Process of Statistics***

Objective 2, Page 1

 *Answer the following while watching the animation.*

1. What is the entire group to be studied called? Population
2. What do we call a person or object that is a member of the population being studied? individual

Objective 2, Page 1 (continued)

1. Give the definition of a sample. Subset of population being studied.
2. What do we call a numerical summary of a sample? A statistic.
3. What do we call a numerical summary of a population? Parameter.
4. Give the definition of descriptive statistics. Sample, extend to the Consists of organizing and summarizing data. It describes data through numerical summaries, tables, and graphs.
5. Give the definition of inferential statistics. Uses results from a sample, extends it to the population and measures reliability of result.
6. In the $100 experiment, what is the population? What is the sample?

Population: complete student body. parameter

Sample: chosen people in student body. statistic

Objective 2, Page 1 (continued)

1. Is the statement an example of descriptive statistics or inferential statistics? Circle the correct answer.
   1. The percent of students in the survey who would return the money to the owner is 78%.

Descriptive statistics right

Inferential statistics

* 1. We are 95% confident that between 74% and 82% of all students would return the money.

Descriptive statistics

Inferential statistics correct

1. Is the given measure a statistic or a parameter? Circle the correct answer.
   1. The percentage of all students on your campus who own a car is 48.2%.

Statistic

Parameter correct

* 1. Suppose a random sample of 100 students is obtained, and from this sample we find that 46% own a car.

Statistic correct

Parameter

Objective 2, Page 7

 *Fill in the following steps while watching the video.*

**The Process of Statistics**

1. **\_identify the problem\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** A researcher must determine the question(s) he or she wants answered. The question(s) must be detailed so that it identifies the population that is to be studied.
2. **\_Collect the Data\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** 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.
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.
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.

Objective 2, Page 8

**Example 1** ***The Process of Statistics: Gun Ownership***

The AP – National Constitution Center conducted a national poll to learn how adult Americans feel existing gun-control laws infringe on the second amendment to the U.S. Constitution.

The following statistical process allowed the researchers to conduct their study.

1. Identify the research objective. Determine % of adult Americans who believe gun control laws infringe on the publics right to bear arms. Population = ADULT AMERICANS
2. Collect the information needed to answer the question posed in (1). Surveyed a SAMPLE of 1007 ADULT AMERICANS. 51% believed existing gun control laws infringe on the publics right bear arms . Cannot survey 200 million americans
3. Describe the data. 51% believe it infringes. DESCRIPTIVE STATISTIC
4. Perform inference 95% chance that all adult americans is between 48% and 54%

***Objective 3: Distinguish between Qualitative and Quantitative Variables***

Objective 3, Page 1

Define the following terms.

1. Qualitative variable: **Qualitative**, or **categorical**, **variables** allow for the classification of individuals based on some attribute or characteristic.
2. Quantitative variable: **Quantitative variables** provide numerical measures of individuals. The values of a quantitative variable can be added or subtracted and provide meaningful results.

Objective 3, Page 2

**Example 2** ***Distinguishing between Qualitative and Quantitative Variables***

Determine whether the following variables are qualitative or quantitative.

1. Gender--qualitative
2. Temperature quantitative
3. Number of days during the past week that a college student studied -quantitative
4. Zip code qualitative

***Objective 4: Distinguish between Discrete and Continuous Variables***

Objective 4, Page 1

Define the following terms.

Variable? Characteristics of the individuals of the population being studied.

1. Discrete variable: A **discrete variable** is a quantitative variable that has either a finite number of possible values or a countable number of possible values. A discrete variable cannot take on every possible value between any two possible values.
2. Continuous variable:

A **continuous variable** is a quantitative variable that has an infinite number of possible values that are not countable. A continuous variable may take on every possible value between any two values.

Objective 4, Page 2

**Example 3** ***Distinguishing between Discrete and Continuous Variables***

Determine whether the quantitative variables are discrete or continuous**. Discrete=count continuous=measure**

1. The number of heads obtained after flipping a coin five times. Discrete
2. The number of cars that arrive at a McDonald’s drive-through between 12:00 P.M. and 1:00 P.M. Discrete
3. The distance a 2011 Toyota Prius can travel in city driving conditions with a full tank of gas. Continuous

Objective 4, Page 4

Define the following terms.

1. Data: The list of observed values for a variable is **data.**
2. Qualitative data: **Qualitative data** are observations corresponding to a qualitative variable.
3. Quantitative data: **Quantitative data** are observations corresponding to a quantitative variable.
4. Discrete data: **Discrete data** are observations corresponding to a discrete variable.
5. Continuous data: **Continuous data** are observations corresponding to a continuous variable.

Objective 4, Page 5

**Example 4** ***Distinguishing between Variables and Data***

The following table presents a group of selected countries and information regarding these countries as of September, 2010.

| **Country** | **Government Type** | **Life Expectancy (years)** | **Population (in millions)** |
| --- | --- | --- | --- |
| 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 |

Identify the individuals, variables, and data.

***Objective 5: Determine the Level of Measurement of a Variable***

Objective 5, Page 1

List the characteristics used to determine what level of measurement a variable is.

1. Nominal: 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
2. Ordinal: if it has the properties of the nominal level of measurement. However, the naming scheme allows for the values of the variable to be arranged in a ranked or specific order

Objective 5, Page 1 (continued)

1. Interval: 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 does not mean the absence of the quantity. Arithmetic operations such as addition and subtraction can be performed on the values of the variable
2. Ration: 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 means the absence of the quantity. Arithmetic operations such as multiplication and division can be performed on the values of the variable.

Objective 5, Page 2

**Example 5 *Determining the Level of Measurement of a Variable***

For each of the following variables, determine the level of measurement.

1. Gender nominal
2. Temperature interval
3. Number of days during the past week that a college student studied ratio
4. Letter grade earned in your statistics class Ordinal

**Section 1.2 Observational Studies versus Designed Experiments**

**Objectives**

1. Distinguish between an Observational Study and a Designed Experiment
2. Explain the Various Types of Observational Studies

***Objective 1: Distinguish between an Observational Study and a Designed Experiment***

Objective 1, Page 1

*Answer the following as you watch the video.*

*1*

Why is the Danish study mentioned in the video an observational study and not a designed experiment? **That is, in an observational study,**

**the researcher simply observes the behavior of the individuals**

**in the study and records the values**

**of the explanatory and response variables.**

2 Why is the “rat” study mentioned in the video a designed experiment and not an observational study **Because the researchers manipulated**

**the value of one explanatory variable, RFR exposure,**

**and controlled other values of explanatory variables,**

**such as temperature, humidity, food,**

**and so on, at fixed values, and there was random assignment,**

**the study is called a designed experiment.**

3.What is the response variable in each study, and what is the explanatory variable? In both studies the goal

was to determine if radio frequencies from cellphones

increase the risk of contracting brain tumors.

**Whether the individual contracted a brain tumor or not**

**was the response variable.**

**For the human study, the explanatory variable**

**of interest was cellphone usage or not.**

**For the rat study, the explanatory variable**

**was radiofrequency radiation, which**

**had three possible levels--**

**no RFR, GSM, or CDMA.**

***In research, we wish to determine***

***how varying an explanatory variable affects the value***

***of a response variable.***

Objective 1, Page 2

*Answer the following after watching the video.*

In research, we wish to determine how varying an explanatory variable affects … ***variable affects the value***

***of a response variable.***

1. What does an observational study measure? Does an observational study attempt to influence the value of the response variable or explanatory variable? Neither

OBJECTIVE 1, PAGE 2 (CONTINUED)

Explain how you would determine if a study is a designed experiment. Whether the explanatory variable values were manipulated or not.

Objective 1, Page 4

*Watch the video and answer the following.*

Why is the influenza study mentioned in the video an observational study and not a designed experiment? Observational. researchers, did not force the seniors to either get a flu shot or not.

But rather it was the seniors who decided that, hey I'm

going to go get a flu shot, or I'm

going to choose not to get a flu shot.

In addition, we as researchers did not

impose any other conditions on the study.

List some changes that could be made to investigate the effectiveness of the flu shot with a designed experiment. We could separate the seniors into two groups and measure more lurking variables and confounding variables. Such as Health, genetics, diet etc…

List some lurking variables in the influenza study. some lurking variables

in the influenza study-- age, health status, mobility

of the senior.

What are some variables (besides getting a flu vaccine) that may play a role in whether one contracts pneumonia or influenza? Genetics, and above lurking variables.

Objective 1, Page 4 (continued)

Define confounding in a study. So again, confounding in a study occurs

when the effect 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

What is a lurking variable? So a lurking variable is an explanatory variable

that was not considered in the study,

but that effects the value of the response variable.

**Do observational studies allow a researcher to claim causality**? No because in observational studies

we are never allowed to make statements of causality.

You cannot say the changes in an explanatory variable cause some

change in a response variable in observational studies.

You can only say that there's an association.

It requires designed experiments to make

statements of causality.

So observational studies do not allow a researcher

Objective 1, Page 7

List some reasons why an observational study would be conducted if causation cannot be claimed. Why would we ever conduct an observational study if we cannot claim causation? Because it is often unethical to conduct a designed experiment, also it has lower cost, greater timeliness, and a broader range of patients

Objective 1, Page 8

Define: Confounding variable confounding in a study occurs

when the effect 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.

Objective 1, Page 9

What is the big difference between lurking variables and confounding variables?

EXAMPLE

In designed experiments, it is possible to have two explanatory variables in a study that are related to each other and related to the response variable. For example, suppose Professor Egner wanted to conduct an experiment in which she compared student success using online homework versus traditional textbook homework. To do the study, she taught her morning statistics class using the online homework and her afternoon class using traditional textbook homework. At the end of the semester, she compared the final exam scores for the online section to the textbook section. If the morning section had higher scores, could Professor Egner conclude that online homework is the cause of higher exam scores? Not necessarily. It is possible that the morning class had students who were more motivated. It is impossible to know whether the outcome was due to the online homework or to the time at which the class was taught. In this sense, we say that the time of day the class is taught is a *confounding variable*.

confounding in a study occurs

when the effect 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.

EXAMPLE

In designed experiments, it is possible to have two explanatory variables in a study that are related to each other and related to the response variable. For example, suppose Professor Egner wanted to conduct an experiment in which she compared student success using online homework versus traditional textbook homework. To do the study, she taught her morning statistics class using the online homework and her afternoon class using traditional textbook homework. At the end of the semester, she compared the final exam scores for the online section to the textbook section. If the morning section had higher scores, could Professor Egner conclude that online homework is the cause of higher exam scores? Not necessarily. It is possible that the morning class had students who were more motivated. It is impossible to know whether the outcome was due to the online homework or to the time at which the class was taught. In this sense, we say that the time of day the class is taught is a *confounding variable*.

**The Difference between Lurking Variables and Confounding Variables**

The big difference between lurking variables and confounding variables is that lurking variables are not considered in the study (for example, we did not consider lifestyle in the pneumonia study) whereas confounding variables are measured in the study (for example, we measured morning versus afternoon classes).

So lurking variables are related to both the explanatory and response variables, and this relation is what creates the apparent association between the explanatory variable and response variable in the study. For example, lifestyle (healthy or not) is associated with the likelihood of getting an influenza shot as well as the likelihood of contracting pneumonia or influenza.

A confounding variable is a variable in a study that does not necessarily have any association with the other explanatory variable but does have an effect on the response variable. Perhaps morning students are more motivated, and this is what led to the higher final exam scores, not the homework delivery system.

The bottom line is that both lurking variables and confounding variables can confound the results of a study, so a researcher should be mindful of their potential existence.

***Objective 2: Explain the Various Types of Observational Studies***

. cross-sectional study.

And this is an observational study that collects information

about individuals at a specific point in time

or over a very short period of time.

Do smokers have a higher resting heart rate than nonsmokers?

You just go out and ask people.

Hey, are you a smoker or are you not a smoker?

That's how you segment your population.

And then, you would measure their heart rate.

This would be a cross-sectional study,

because it's over a very short period of time.

18. A case control study.

These studies are retrospective.

Retrospective basically means that you look back in time.

Case control studies look back in time

or require the researcher or to look at existing record, which

is also looking back in time.

In case control studies, individuals

that have certain characteristics

are matched with those that do not.

That's why it's case control, because with this smoker thing

we might say, well, how many packs of cigarettes

have you smoked on average each week for the past 10 years.

And we would match those folks with people

with similar demographics, characteristics

that did not smoke.

And then, we might compare the rate of lung cancer

in the two groups.

A case control study.

These studies are retrospective.

Retrospective basically means that you look back in time.

Case control studies look back in time

or require the researcher or to look at existing record, which

is also looking back in time.

In case control studies, individuals

that have certain characteristics

are matched with those that do not.

That's why it's case control, because with this smoker thing

we might say, well, how many packs of cigarettes

have you smoked on average each week for the past 10 years.

And we would match those folks with people

with similar demographics, characteristics

that did not smoke.

And then, we might compare the rate of lung cancer

in the two groups.

19. The negative thing about case control studies

is there's a memory issue here.

People might not remember how many packs of cigarettes

they smoked on average for the past 10 years.

And therefore, you might get bogus information out

of the study.

Or records might not exist or might be somewhat flawed,

which impacts the study.

But case control studies, none the less,

are better than cross-sectional studies,

because you have a matching going on.

You have a control group to compare

against your sort of research group, nonsmokers

versus the smokers.

But the negative, again, is that, looking back in time,

you often get erroneous information.

21. A cohort study first identifies a group of individuals

to participate in the study.

This is called your cohort.

And then, the cohort is observed over a period of time.

Over this period of time, characteristics

about individuals are recorded.

So these are called prospective studies,

because they are done going forward.

So basically, what I would do is something along these lines.

I would identify 5,000 individuals

and literally follow them over time, periodically

recording information about the individuals.

22. The advantage here is that you don't

have to rely on folks looking back in time.

And you don't have to rely on existing records,

because you are collecting the information over time.

The famous cohort study that exists

is the Framingham Heart Study, which

has been going on for years.

And in fact, is now in new generations

and because of this Framingham Heart cohort study,

a lot of new ideas and breakthroughs in medicine

have occurred.

23. The negative about these studies is

that they are extremely time intensive and labor

intensive and, therefore, usually very expensive.

Objective 2, Page 2

**Example 1** ***What Type of Study?***

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.

1. Researchers wanted to assess the long-term psychological effects of 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.)
2. Xylitol has proven effective in preventing dental carries (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 carries prevention. BMC Oral Health (5)6.)
3. 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. (Source: Swanson KA, Andersen R, Gelberg L (2003) Patient satisfaction for homeless women. Journal of Women’s Health (12)7.)
4. 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)

Type of Study?

A. Observational- Case Control

B. Experimental

C. Observational- Cross Section

D. Observational- Cohort

Objective 2, Page 3

It is not always possible to conduct an experiment. Explain why we could not conduct an experiment to investigate the perceived link between high tension wires and leukemia (on humans**). It is not ETHICAL to conduct some experiments.**

**me Concluding Remarks about Observational Studies Versus Designed Experiments**

Is a designed experiment superior to an observational study? Not necessarily.

* Because cross-sectional and case-control observational studies are relatively inexpensive, they allow researchers to explore possible associations prior to undertaking large cohort studies or designed experiments.
* It is not always possible to conduct an experiment. For example, we could not conduct an experiment to investigate the perceived link between high-tension wires and leukemia (on humans). Do you see why?

25. Have you heard this saying?

*"There is no point in reinventing the wheel."*

Here is how it applies to statistics: There is no sense expending energy to obtain data that already exist. If a researcher wants to conduct a study and an appropriate data set exists, it would be silly to collect the data from scratch.

For example, various federal agencies regularly collect data that are available to the public. Some of these agencies include the Centers for Disease Control and Prevention ([www.cdc.gov](http://www.cdc.gov/)), the Internal Revenue Service ([www.irs.gov](http://www.irs.gov/)), and the Department of Justice (<http://fjsrc.urban.org/index.cfm>). Another useful source of data is the General Social Survey (GSS), [www.gss.norc.org](http://www.gss.norc.org/), administered by the University of Chicago. This survey regularly asks "demographic and attitudinal questions" of individuals around the country.

26.   
A **census** is a list of individuals in a population along with certain characteristics of each individual.

27. The United States conducts a census every 10 years to learn the demographic makeup of the United States. Everyone whose primary residence is within the U.S. borders must fill out a questionnaire packet. The cost of obtaining the census in 2010 was approximately $5.4 billion; about 635,000 temporary workers were hired to assist in collecting the data.

Why is the U.S. Census so important? **The results of the census are used to determine the number of representatives in each state in the House of Representatives, boundaries of congressional districts, distribution of funds for government programs (such as Medicaid), and planning for the construction of roads and schools.** The first U.S. Census was conducted in 1790 under the direction of Thomas Jefferson. It is a constitutional mandate that a census be conducted every 10 years.

Is the United States successful in obtaining a census? Not entirely. Some individuals go uncounted due to illiteracy, language issues, and homelessness. Given the political stakes that are based on the census, politicians often consider how to count these individuals. Statisticians have offered solutions to the counting problem. If you wish, go to [www.census.gov](http://www.census.gov/); in the search box, type *count homeless*. You will find many articles on the U.S. Census Bureau's attempt to count the homeless. The bottom line is that even census data has flaws.

**Section 1.3 Simple Random Sampling**

**Objective**

* 1. Obtain a Simple Random Sample

Introduction, Page 1

Observational studies can be conducted by administering a survey. When administering a survey, the researcher must first identify the population that is to be targeted.

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

For the results of a survey to be reliable, the characteristics of the individuals in the sample must be representative of the characteristics of the individuals in the population. The key to obtaining a sample representative of a population is to let *chance* or *randomness,* rather than convenience, play a role in dictating which individuals are in the sample. **If convenience is used to obtain a sample, the results of the survey are meaningless.**

For the results of a survey to be reliable, the characteristics of the individuals in the sample must be representative of the characteristics of the individuals in the population.

The key to obtaining a sample representative of a population is to let chance or randomness play a role in dictating which individuals are in the sample, rather than convenience.

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

Introduction, Page 2

2.Why are the survey results from the sample taken outside Fenway Park not likely to be reliable?  Clearly, the individuals in the sample do not accurately reflect the makeup of the entire population, CONVENIENCE.

3.Why are the results of a survey of students in your statistics class likely to be misleading when trying to determine what proportion of students on your campus work? The convenient sample is not representative of the population, which means that any results reported from your survey are misleading. Does your class mirror the gender, grade, or day/ evening of the entire campus, probably not.

Introduction, Page 3

List the four basic sampling techniques.

1. Simple random sampling
2. Stratified sampling
3. Systematic sampling
4. Cluster sampling

***Objective 1: Obtain a Simple Random Sample***

Objective 1, Page 1

1. What is a simple random sample? 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 equal chance of occurring. The sample is then called a **simple random sample.**

The number of individuals in the sample is always less than the number of individuals in the population.

Objective 1, Page 2

**Example 1** ***Illustrating Simple Random Sampling***

Sophie 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. Sophie decides to randomly select three of her six friends to attend the concert.

1. List all possible samples of size *n* = 3 from the population of size *N* = 6. Once an individual is chosen, he/she cannot be chosen again. 20 total
2. Comment on the likelihood of the sample containing Michael, Kevin, and Marissa. 1 in 20

Objective 1, Page 5

How do we select the individuals in a simple random sample?

Typically, each individual in the population is assigned a unique number between 1 and *N*, where *N* is the size of the population. Then *n* distinct random numbers are selected, where *n* is the size of the sample. To number the individuals in the population, we need a frame**–** a list of all the individuals within the population. ***How do we select the individuals in a simple random sample?***

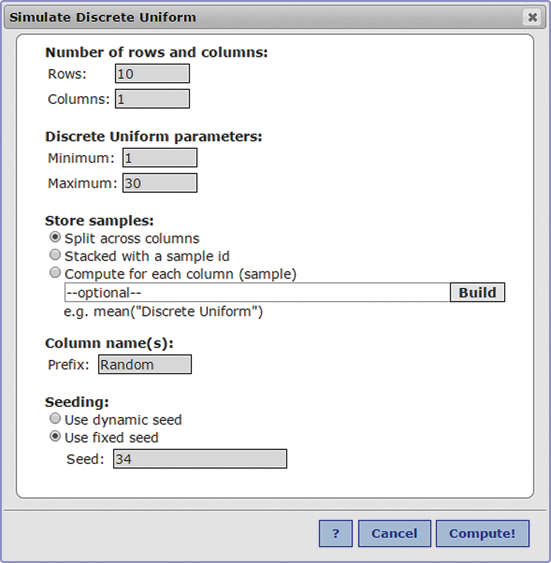
We could write the names of the individuals in the population on different pieces of paper and then select names from a hat. Often, however, the size of the population is so large that performing simple random sampling in this fashion is not practical.

Typically, each individual in the population is assigned a unique number between 1 and N, where N is the size of the population. Then n distinct random numbers are selected, where n is the size of the sample.

**To number the individuals in the population, we need a frame—a list of all the individuals within the population.**

#### **Obtaining a Simple Random Sample**

1. Select **Data**, highlight **Simulate**, and then highlight **Discrete Uniform**.
2. Fill in the window with the appropriate values. To obtain a simple random sample of size n=5 from a population of size N=30, enter the values shown in the figure. The reason we generate 10 rows of data (instead of 5) is in case any of the random numbers repeat. Click Compute! and the random numbers will appear in the spreadsheet. *Note*: You could also select the single dynamic seed radio button to set the seed.



Objective 1, Page 6

*Answer the following after watching the animation.*

1. What is the frame in this animation? Students in class

Explain why a second sample of 5 students will most likely be different than the first sample of 5 students? **The individuals differ from sample to sample**

**because chance is used to select the individuals.**

Explain why inferences based on samples vary. **Inferences based on samples will vary**

**because the individuals in the samples vary.**

Objective 1, Page 8

**Example 2 *Obtaining a Simple Random Sample***

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

**TABLE 3**

* + 1. ABC Electric
    2. Brassil Construction
    3. Bridal Zone
    4. Casey's Glass House
    5. Chicago Locksmith
    6. DeSoto Painting
    7. Dino Jump
    8. Euro Car Care
    9. Farrell's Antiques
    10. First Fifth Bank
    11. Fox Studios
    12. Haynes Hauling
    13. House of Hair
    14. John's Bakery
    15. Logistics Management, Inc.
    16. Lucky Larry's Bistro
    17. Moe's Exterminating
    18. Nick's Tavern
    19. Orion Bowling
    20. Precise Plumbing
    21. R&Q Realty
    22. Ritter Engineering
    23. Simplex Forms
    24. Spruce Landscaping
    25. Thors, Robert DPS
    26. Travel Zone
    27. Ultimate Electric
    28. Venetian Gardens Restaurant
    29. Walker Insurance

Worldwide Wire

Suppose Professor Cummings wishes

to estimate the average travel time to school for his class.

Rather than surveying each of the 33 students enrolled,

he decides to obtain a simple random sample of 5 students.

Professor Cummings already has a frame.

Remember, a "frame" is a list of all the individuals

in the population of interest.

In this case, the frame is the list

of students enrolled in the class.

To obtain a simple random sample,

Professor Cummings first assigns a unique number

to each student in the class.

So the first student in the class

is assigned number 1, the second student is assigned the number

2, and so on until we reach the last student who

is assigned the number 33.

5 different numbers will be randomly selected.

The students corresponding to these numbers

are the individuals in the sample.

This is sampling without replacement,

which means that an individual who is selected to be

in the sample from the population cannot be selected

again.

On the screen, we have our 33 students--

numbered 1 through 33.

To get the random numbers used to generate

our simple random sample, we could draw the numbers out

of an urn.

We select "Sample Students," and we draw our 5 numbers.

So the students that are in our sample

are student 8, student 24, student 14, student 15,

and student 32.

Student 8, student 24, student 14, student 15, and student 32.

We find that the average time of their commute is 25 minutes.

Let's obtain a second random sample of size 5

from our student population.

The procedure is the same.

We're going to randomly select 5 unique numbers from an urn.

The individuals corresponding to the numbers are in the sample.

Student 15, student 29, student 1, student 17, and student 6.

We then ask these students to report their travel time

to school.

19 minutes, 18 minutes, 4 minutes, 3 minutes,

and 10 minutes.

We get an average travel time to school of 10.8 minutes.

Notice that the individuals in our first sample

are different from the individuals

in the second sample.

Our first sample had Megan, Uri, Adam, Suman, and Keith.

Our second sample had Alizandro, Matt, Adam, Crystal, and Megan.

For this reason, each sample results

in different descriptive statistics.

The first sample had an average commute time of 25 minutes.

The second sample had an average commute time of 10.8 minutes.

Therefore, any inference based on each sample

may result in different conclusions regarding

the population.

This is the very nature of statistics.

Inferences based on samples will vary

because the individuals in the samples vary.

The individuals differ from sample to sample

because chance is used to select the individuals.

Of course, for large populations--

such as all the students enrolled

in a particular college or university--

the approach just taken in using an urn

would be difficult to use.

Imagine a school with 15,000 students.

It would be quite a large urn that

would require 15,000 balls.

Instead, in practice, random number generators--

such as those found on TI graphing calculators

or StatCrunch-- are used to obtain the individuals

in the simple random sample.

**Section 1.4 Other Effective Sampling Methods**

**Objectives**

* 1. Obtain a Stratified Sample
  2. Obtain a Systematic Sample
  3. Obtain a Cluster Sample

Introduction, Page 1

What is the goal of sampling? **The goal of sampling is to collect as much information as possible about the population at the least cost. Cost includes monetary outlays, time, and other resources. With this goal in mind, we may find it advantageous to use sampling techniques other than simple random sampling.**

***Objective 1: Obtain a Stratified Sample***

Objective 1, Page 1

Explain how to obtain a stratified sample. **STRATIFIED SAMPLE**

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

An advantage of stratified sampling over simple random sampling is that it may allow fewer individuals to be surveyed while it obtains the same or more information. This result occurs because individuals within each subgroup have similar characteristics, so opinions within the group are not as likely to vary much from one individual to the next. In addition, a stratified sample guarantees that each stratum is represented in the sample

In this example, we'll go over the procedure

for obtaining a stratified sample.

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 or commuting students, and staff-- including

faculty-- so that he can obtain a stratified sample.

Suppose there are 6,204 resident students, 13,304

nonresident students, and 2,401 staff

members for a total of 21,909 individuals in the population.

The resident students make up approximately 28%

of the population, the non-resident students make up

approximately 61% of the population,

and staff makes up approximately 11% of the population.

We want a stratified sample to represent this percentage

breakdown so that it looks just like the population, only

smaller.

The president wants to obtain a sample of size 100

with the number of individual selected

from each stratum weighted by the population size.

So we'll multiply the desired percentage

by the sample size of 100.

That would provide us with 28 resident students

in our sample, 61 non-resident students, and 11 staff members.

We simply multiply the percentage

by the sample size for each group.

To obtain the stratified sample, we'll

construct a simple random sample within each group.

We'll select 28 of the 6,204 resident students,

61 out of the 13,304 nonresident students,

and 11 out of the 2,401 staff members.

If you're going to do this with technology-- and we

will-- be careful not to use the same seed for all the groups

in the stratified sample because we

want the simple random samples within each stratum

to be independent of each other.

STEPS

DATA

SIMULATE

DISCRETE UNIFORM

**Again, I need 28 values, but I'm going to ask for 5 more**

**than that in case there are duplicates.**

**So 33 rows, 1 column, the minimum number**

**is 1, the maximum number 6,204, and I'm**

**going to use the dynamic seed.**

**Press compute.**

**So I'm going to use StatCrunch to select the random numbers.**

**I'm going to begin with the resident students.**

**We needed 28 students out of 6,204 students.**

**To select those numbers, press on Data, Simulate, Discrete**

**Uniform.**

**Again, I need 28 values, but I'm going to ask for 5 more**

**than that in case there are duplicates.**

**So 33 rows, 1 column, the minimum number**

**is 1, the maximum number 6,204, and I'm**

**going to use the dynamic seed.**

**Press compute.**

**Let's go ahead and generate the 61 non-resident students.**

**Data, Simulate, Discrete Uniform.**

**I needed 61.**

**I'm going to ask for 66, again, in case there are repeats.**

**The minimum value is 1.**

**The maximum value is 13,304.**

**I'll use another dynamic seed.**

**Press Compute.**

**Now to generate the 11 staff.**

**Data, Simulate, Discrete Uniform.**

**I need 11, so I'll ask for 16 in one column.**

**The minimum value is 1.**

**The maximum value, 2,401.**

**I'm going to save myself some grief and where it says Prefix,**

**I'm going to put the column name.**

**This one's going to be called Staff,**

**and I won't have to change it later.**

**Again, I'll use a dynamic seed, let the computer select it.**

**Press Compute, and that column,**

And I'm going to get rid of this header

and replace it with Resident.

Now I will only use the first 28 of those values that

do not repeat.

I'll check them for repeats in a moment.

Let's go ahead and generate the 61 non-resident students.

Data, Simulate, Discrete Uniform.

I needed 61.

I'm going to ask for 66, again, in case there are repeats.

The minimum value is 1.

The maximum value is 13,304.

I'll use another dynamic seed.

Press Compute.

And I'm going to change that header to Non-resident.

I'll check that for repeats in a bit as well.

We'll only use the first 61 that are non-repeating.

Now to generate the 11 staff.

Data, Simulate, Discrete Uniform.

I need 11, so I'll ask for 16 in one column.

The minimum value is 1.

The maximum value, 2,401.

I'm going to save myself some grief and where it says Prefix,

I'm going to put the column name.

This one's going to be called Staff,

and I won't have to change it later.

Again, I'll use a dynamic seed, let the computer select it.

Press Compute, and that column, Staff 1-- I'll just get rid

of the 1-- has been added.

And I'm only going to use the first 11 that do not repeat.

It looks like the first 11 there are clean.

OK, here are the 33 random numbers

we generated for resident students.

We needed 28, which means that we're supposed to end here.

And we've scoured the first 28 looking for any repeats,

and in this case, there are none,

so we can discard the 5 extra values that I requested.

And there are the 28 resident students

that will be included in the study.

Now your 28 and my 28 will be different

because they're selected at random.

Let's look at the non-resident students.

Selected 66 values.

I only need the first 61.

And in the first 61, I go looking for repeats,

and there actually are repeats.

12,731 was elected twice, so I can discard the second one,

and my first 61 values-- that was 61.

Add one more and discard the rest like this.

Now again, your 61 will be different than my 61

because we selected them randomly.

Finally, let's take a look at the staff.

Here are the 16 values I selected.

I only need 11 of them.

So as I look at those first 11 staff members,

there are no repeats, so I can get rid of the extra 5

I asked for.

And there are the 11 staff members

that will be included in the study.

So we used the first 28 non-repeating values

that we found for the resident, the first 61

non-repeating values for the non-residents,

and the first 11 non-repeating values for the staff.

You could have done this using Microsoft Excel,

some other software package.

You also could have done this with a TI calculator.

Just keep going until you get your 28 unique residents, 61

unique non-residents, and so on.

An advantage of stratified sampling

over simple random sampling is that the researcher

can determine characteristics within each strata.

This allows an analysis to be performed on each stratum

to see if any significant differences among them exist.

For example, we could analyze the data obtained in example 1

to see if there's a difference in the opinions of students

versus staff.

Objective 1, Page 2

**Example 1** ***Obtaining a Stratified Sample***

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 6,204 resident students, 13,304 nonresident students, and 2,401 staff, for a total of 21,909 individuals in the population. What percent of the DePaul community is made up of each group?

The president wants to obtain a sample of size 100, with the number of individuals selected from each stratum weighted by the population size. How many individuals should be selected from each stratum?

To obtain the stratified sample, construct a simple random sample within each group.

***Objective 2: Obtain a Systematic Sample***

Objective 2, Page 1

Explain how to obtain a systematic sample. **SYSTEMATIC SAMPLE**

In both simple random sampling and stratified sampling, a [frame](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_1_c7c47617-8634-539a-37e1-6d2a73e7cae4) must exist. Therefore, these sampling techniques require some preliminary work before the sample can be found. A sampling technique that does not require a frame is *systematic sampling*.

A **systematic sample** is obtained by selecting every kth individual from the population. The first individual selected corresponds to a number between 1 and k.

Because systematic sampling does not require a frame, it is a useful technique when you cannot gather a list of the individuals in the population. Also, systematic samples typically provide more information for a given cost than does simple random sampling. In addition, systematic sampling is easier to employ; so there is less likelihood of interviewer error occurring, such as selecting the wrong individual to be surveyed.

**Note:** Because systematic sampling does not require a frame, it is a useful technique when you cannot gather a list of the individuals in the population.

Objective 2, Page 2

**Example 2 *Obtaining a Systematic Sample without a Frame***

The manager of Kroger Food Stores wants to measure the satisfaction of the store’s customers. Design a sampling technique that can be used to obtain a sample of 40 customers.

STATCRUNCH

APPLET

RANDOM NUMBERS

7th client

1

7

Compute

3,7,14…….1 less than sample of 40 3 + 39(7) times every 7th

Objective 2, Page 4

*Answer the following after watching the video.*

1. that can result from choosing a value of *k* that is too small

If the size of the population is unknown,

there's no mathematical way to determine k

for systematic sampling.

The value of k must be small enough

to achieve our desired sample size.

But it also must be large enough to obtain a sample that's

representative of the population.

For example 2, suppose that k was equal to 30.

We said that we would start with customer 3,

and then add 30 to that, and 30 to that,

until we reach our 40th customer, which we could find

by multiplying 39 times 30 added to the original 3,

that's customer number 1,173.

If Kroger does not have 1,173 shoppers,

the desire sample size will not be achieved.

Now suppose that k equals 4, starting again at 3,

the second person selected would be customer 7, and then

customer 11, until we reach our 40th customer selected,

that's 39 times 4 plus 3, that's customer number 159.

The 159 shopper might leave the store at 3:00 in the afternoon,

so our survey would not include any of the evening shoppers.

An estimate of the size of the population

would help to determine an appropriate value of k.

#### W **Choosing a Value for**k

When using systematic sampling, how would we select the value of k?

If the size of the population is unknown, there is no mathematical way to determine k. The value of k must be small enough to achieve our desired sample size and large enough to obtain a sample that is representative of the population. The following video illustrates the importance of k

What can result from choosing a value of *k* that is too large? The value of k must be small enough to achieve our desired sample size

Objective 2, Page 5 

 *Answer the following after watching the second video after Example 2.*

Explain how to determine the value of *k* if the population size *N* is known.

Now we'll take a look at how to determine k

when the population size n is known.

So if possible, approximate the population size n.

Then determine the sample size desired-- lowercase n.

Divide the population size by the sample size

and round down to the nearest integer.

This value is k.

So let's suppose that the population size was 20,325.

And we desire a sample size of n equals 100.

If we divide 20,325 by 100, that's equal to 203.25.

And rounding this down-- that gives us a value of k of 203.

Now let's further suppose that we

start with the 90th individual.

So our sample would include the 90th individual.

Then I add 203.

That's 293.

I add 203 again.

That's 496.

And we keep going until we reach our desired sample size.

The last value will be 1 less than the sample

size-- 99-- times the value of k-- 203--

added to the starting value of 90.

And that equals 20,187.

So the 20,187th individual will be the last individual

in the survey.

Objective 2, Page 7

1. List the five steps in obtaining a systematic sample.

#### **Steps in Systematic Sampling**

1. If possible, approximate the population size, N.
2. Determine the sample size, n.
3. Compute Nn and round down to the nearest integer. This value is k.
4. Randomly select a number between 1 and k. Call this number p.
5. The sample will consist of the following individuals:
6. p,p+k,p+2k,...,p+(n−1)k

***Objective 3: Obtain a Cluster Sample***

Objective 3, Page 1

A fourth sampling method is called *cluster sampling*. Like the previous three sampling methods, this method has benefits under certain circumstances.

**DEFINITION**

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

Suppose a school administrator wants to learn the characteristics of students enrolled in online classes. Rather than obtaining a simple random sample based on the frame of all students enrolled in online classes, the administrator treats each online class as a cluster and then finds a simple random sample of these clusters. The administrator then surveys *all* students in the selected clusters. This reduces the number of classes that get surveyed

Objective 3, Page 2

In this example, we'll explore the process

of obtaining a cluster sample.

A sociologist wants to gather data regarding

household income within the city of Boston.

Obtain a sample using cluster sampling.

The city of Boston can be set up so

that each city block is a cluster.

Recall that a cluster is a geographic location.

Once the city blocks have been identified,

we obtain a simple random sample of the city blocks,

and then we survey all households

on the block selected.

Let's suppose there are 10,493 city blocks in Boston.

First, the sociologist has to number these blocks from 1 up

through 10,493.

Suppose the sociologist only has enough time and money

to survey 20 clusters.

We'll randomly select 20 numbers between 1 and 10,493

to tell us which clusters, or city blocks,

will be included in the survey.

We have many options to select 20 blocks from the 10,493

blocks.

**I'm going to use StatCrunch.**

**Press Data, Simulate, Discrete Uniform.**

**I only need 20 values, but I'm going**

**to ask for 25 in case there are repeats.**

**I'm going to put them all in one column.**

**The minimum value is 1.**

**The maximum value is 10,493.**

**I'm going to use a dynamic seed.**

**Let the computer pick it, and press compute.**

**And my values are here in the first column.**

**I'm going to check for repeats, and if there are none,**

**I'll use the first 20 values.**

**So here are the 25 random numbers that we selected.**

**I only need 20, which will end above the line.**

**I want to take a look at those first 20**

**and see if there are any repeats.**

**And since there are not, I can discard the extra 5**

**random numbers I generated.**

**So these are the 20 blocks that will be included in the study.**

There are a couple of advantages of cluster

sampling we should discuss.

First, cluster sampling reduces the travel time

that would likely be required with stratified

sampling or simple random sampling in that many

of the households will be clustered together

rather than driving across the city.

Second, there's no need to obtain

a frame of all the households.

All we really need is one frame that

provides information regarding city blocks.

**Example 3 *Obtaining a Cluster Sample***

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

Objective 3, Page 3

1. If the clusters have homogeneous individuals, is it better to have more clusters with fewer individuals in each cluster or fewer clusters with more individuals in each cluster?
2. If the clusters have heterogeneous individuals, is it better to have more clusters with fewer individuals in each cluster or fewer clusters with more individuals in each cluster?

**Issues to Consider in Cluster Sampling**

The following questions arise in cluster sampling:

* How do I cluster the population?
* How many clusters do I sample?
* How many individuals should be in each cluster?

First, we must determine whether the individuals within the proposed cluster are *homogeneous* (similar individuals) or *heterogeneous* (dissimilar individuals).

In Example 3, city blocks tend to have similar households. Survey responses from houses on the same city block are likely to be similar. This results in duplicate information. We conclude that **if the clusters have homogeneous individuals, it is better to have more clusters with fewer individuals in each cluster.**

1. If the clusters have homogeneous individuals, is it better to have more clusters with fewer individuals in each cluster or fewer clusters with more individuals in each cluster?

**If the clusters have homogeneous individuals, it is better to have more clusters with fewer individuals in each cluster.**

1. If the clusters have heterogeneous individuals, is it better to have more clusters with fewer individuals in each cluster or fewer clusters with more individuals in each cluster?

What if the cluster is heterogeneous? Under this circumstance, the heterogeneity of the cluster likely resembles the heterogeneity of the population. In other words, each cluster is a scaled-down representation of the overall population.

For example, a quality control manager might use shipping boxes that contain 100 lightbulbs as a cluster. The manager does this because the rate of defects within the cluster resembles the rate of defects in the population, assuming that the bulbs are randomly placed in the box. Thus,

**When each cluster is heterogeneous, fewer clusters with more individuals in each cluster are appropriate.**

Objective 3, Page 5

1. Define: Convenience sampling

Objective 3, Page 6

**Note:** The most popular convenience samples are those in which the individuals in the sample are self-selected**,** meaning the individuals themselves decide to participate in the survey. Self-selected surveys are also called voluntary response samples.

Objective 3, Page 7

1. Define: Multistage sampling
2. List the two stages Nielsen Media Research uses to investigate TV viewing habits.

Objective 3, Page 8

1. How many stages does the Census Bureau use for the Current Population Survey? What are those stages?

Objective 3, Page 9

Researchers need to know how many individuals they must survey to draw conclusions about the population within some predetermined margin of error. They must find a balance between the reliability of the results and the cost of obtaining these results. The bottom line is that time and money determine the level of confidence researchers will place on the conclusions drawn from the sample data. The more time and money researchers have available, the more accurate the results of the statistical inference.

Objective 3, Page 10

 *Watch the animation for a summary of simple random sampling, systematic sampling, stratified sampling, and cluster sampling.*

**Convenience Sampling**

In the four sampling techniques just presented (simple random sampling, stratified sampling, systematic sampling, and cluster sampling), the individuals are selected randomly. Often, however, inappropriate sampling methods are used in which the individuals are *not* randomly selected.

Have you ever been stopped in the mall by someone holding a clipboard? These folks are responsible for gathering information, but their methods of data collection are inappropriate, and the results of their analysis are suspect because they collect data using a *convenience sample*.

**DEFINITION**

In a **convenience sample,** the individuals are easily obtained and not based on randomness.

**Examples of Convenience Samples**

The most popular convenience samples are those in which the individuals in the sample are **self-selected**, meaning the individuals themselves decide to participate in the survey. Self-selected surveys are also called **voluntary response** samples. One example of self-selected sampling is phone-in polling—a radio personality will ask his or her listeners to phone or text the station to submit their opinions. Another example is the use of the Internet to conduct surveys. For example, a TV news show will present a story regarding a certain topic and ask its viewers to "tell us what you think" by completing an online questionnaire or tweeting an opinion with a hashtag.

Both of these samples are poor designs because the individuals who decide to be in the sample generally have strong opinions about the topic. A more typical individual in the population will not bother phoning, texting, or tweeting to complete a survey. Any inference made regarding the population from this type of sample should be made with extreme caution.

**Multistage Sampling**

**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, through a People Meter, monitors the television programs these households are watching. The meter is an electronic box connected to 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.

**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 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.

**An Example of Multistage Sampling**

Consider the sample used by the Census Bureau for the Current Population Survey. This survey requires five stages of sampling:

**Stage 1:** Stratified sample  
  
**Stage 2:** Cluster sample  
  
**Stage 3:**Stratified sample  
  
**Stage 4:** Cluster sample  
  
**Stage 5:** Systematic sample

This survey is very important because it is used to obtain demographic estimates of the United States in noncensus years. Details about the Census Bureau's sampling methods can be found in *The Current Population Survey: Design and Methodology*, Technical Paper No. 40

**Sample Size Considerations**

Throughout our discussion of sampling, we did not mention how to determine the sample size. Researchers need to know how many individuals they must survey to draw conclusions about the population within some predetermined margin of error.

Researchers must find a balance between the reliability of the results and the cost of obtaining these results. Time and money determine the level of confidence researchers will place on the conclusions drawn from the sample data. The more time and money researchers have available, the more accurate the results of the statistical inference will be.

Later in the course, we will discuss techniques for determining the sample size required to estimate characteristics regarding the population within some margin of error. (For a detailed discussion of sample size considerations, consult a text on sampling techniques such as *Elements of Sampling Theory and Methods* by Z. Govindarajulu, Pearson, 1999.)

VIDEO -SUMMARY OF SAMPLING TECHNIQUES

STUDENT: Can you help me distinguish the four

different sampling methods?

**Simple random sampling** is like drawing names from a hat.

We use a frame to assign numbers to each individual

in the population.

Then random numbers are generated

to identify the individuals in the sample.

For example, suppose we have a population

of all introductory statistics students with 100 individuals.

We would like to know the average GPA

of an introductory statistics student,

so decide to obtain a simple random sample of 5 students.

We number the individuals from 1 to 100

and randomly generate 5 unique numbers to obtain the sample.

So in this particular simple random sample,

we obtained student 9, 74, 90, 61, and 63,

and obtained an average GPA of 3.312.

**systematic sampling**

INSTRUCTOR: With systematic sampling,

we sample every kth individual.

Again, consider the 100 introductory statistics

students.

Suppose we want a sample of size 5 from this population.

Because the population size is 100, we divide 100 by 5

and obtain 20.

Therefore, randomly select a number between 1 and 20

and survey every 20th individual in the population.

The key with systematic sampling is

that the individuals in the population

are arranged in some order.

Even if you do not have a frame, a systematic sample

could be obtained, such as selecting every 10th can

from an assembly line.

So in this particular sample, we randomly selected the number 7

between 1 and 20, and so we survey

the 7th, 27th, 47th, 67th, and 87th student,

and obtain a mean GPA of 3.476.

**stratified sampling**,

we divide the population into non-overlapping homogeneous

groups.

That is, each group must have some common characteristic.

For example, suppose we divide the class

into on-campus students in red and off-campus students

in blue.

Then within each of these strata,

we might obtain a simple random sample of 5 students.

For example, for the on-campus students,

we selected student 25, 91, 61, 65, and 63.

For the off-campus students, we selected students 64, 34, 49,

6, and 30.

The mean GPA of the on-campus student is 2.941.

The mean GPA of the off-campus student is 2.719.

**cluster sampling**

INSTRUCTOR: Cluster sampling is obtained

by selecting all the individuals within a collection

of individuals.

Again, consider our classroom of 100 statistics students.

We could treat each row students as a cluster.

Then we might randomly select two rows.

Each student in the selected rows is then surveyed.

So here we randomly selected row five, and each student in row

five is surveyed for their GPA.

We also randomly selected row four.

And again, each student in row four

this surveyed to determine their GPA.

We find that the mean GPA of the 20 students surveyed is 3.019.

**How does stratified sampling**

**differ from cluster sampling?**

:

INSTRUCTOR: With stratified sampling,

we divide the population into non-overlapping groups,

such as freshmen, sophomores, juniors, seniors.

Then within each strata, we obtain a simple random sample.

This is great if we wish to compare

the value of the variable, such as GPA,

between the various strata.

Do freshman have a different GPA from sophomores?

With cluster sampling, the population is also divided,

but we randomly select clusters and then survey all individuals

within the cluster.

For example, suppose we wanted to find out

why students enroll in 8:00 AM courses on Monday morning.

You might divide the population of all 8:00 AM Monday courses

into clusters.

Then we obtain a simple random sample of four courses,

and survey all students in the four selected

**Section 1.5 Bias in Sampling**

**Objective**

1. Explain the Sources of Bias in Sampling

***Objective 1: Explain the Sources of Bias in Sampling***

Objective 1, Page 1

1. Define: Bias If the results of the sample are not representative of the population, then the sample has **bias**.

List the three sources of bias in sampling: There are **three** sources of bias in sampling:

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

Objective 1, Page 2

 *Answer the following after watching the video.*

What is sampling bias? **Sampling bias**

means that the technique used

to obtain the samples individuals

tends to favor one part of the population over another.

Any convenient sampling has sampling bias

because the individuals are not chosen through a random sample.

**Sampling bias** also results due to

under coverage, which occurs when

the proportion of one segment of the population

is lower in a sample then it is in the population.

**Undercoverage** can result if the frame used to obtain the sample

is incomplete or not representative

of the population.

Some frames, such as the list of all registered voters,

may seem easy to obtain.

But even this frame may be incomplete since people

who recently registered to vote may not

be on the publish list of registered voters.

Sampling bias can lead to incorrect predictions.

For example, the magazine Literary Digest

predicted that Alfred M Landon would defeat Franklin D

Roosevelt in the 1936 presidential election.

**The Literary Digest conducted a poll**

**based on a list of its subscribers**, telephone

directories, and automobile owners.

On the basis of the results, the Literary Digest

predicted that Landon would win with 57% of the popular vote.

However, Roosevelt won the election with about 62%

of the popular vote.

**This election took place during the height**

**of the Great Depression.**

**In 1936, most subscribers to the magazine, households**

**with telephones, and automobile owners**

**were Republican-- the party of Landon.**

Therefore, the choice of the frame

used to conduct the survey led to an incorrect prediction

due to sampling bias.

**Essentially, there was an undercoverage of Democrats.**

It is difficult to gain access to a complete list

of the individuals in a population.

For example, public opinion polls

often use random telephone surveys,

which implies that the frame is all households with telephones.

This method of sampling excludes households

without telephones, as well as homeless people.

If these people differ in some way

from people with telephones or homes,

the results of the sample may not be valid.

The federal government prohibits the random calling

of cellular telephones.

However, a study done by Pew Research Center

suggests that excluding cellphone only homes

from telephone surveys does not significantly

affect the results.

1. What is sampling bias?

**means that the technique used**

**to obtain the samples individuals**

**tends to favor one part of the population over another.**

**Does a convenience sample have sampling bias?**

**Probably, Literary Digest example……Landon over Roosevelt**

What is under coverage? **Undercoverage can result if the frame used to obtain the sample**

**is incomplete or not representative**

**of the population.**

Objective 1, Page 3

 *Answer the following after watching the video.*

When does nonresponse bias exist? **Nonresponse Bias**

Non-response bias exists when individuals selected

to be in the sample who do not respond to the survey

have different opinions from those who do.

List two causes of nonresponse bias. **Non-response can occur because individuals selected**

**for the sample do not wish to respond**

**or the interviewer was unable to contact them.**

All surveys will suffer from non-response.

The federal government's current population survey

has a response rate of about 92%,

but it varies depending on the age of the individual.

For example, the response rate for 20- to 29-year-olds is 85%,

and for individuals 70 and older, it's 99%.

Response rates in random digit dialing telephone surveys

are typically around 70%, email survey response rates

hover around 40%, and mail surveys

can have response rates as high as 60%.

Non-response bias can be controlled using callbacks.

For example, if a mail questionnaire was not returned,

a callback might mean phoning the individual

to conduct the survey.

If an individual is not at home, a callback

might mean returning to the home at other times in the day.

Another method to improve non-response

is using rewards such as cash payments

for completing a questionnaire or incentives such as a cover

letter that states that the responses to the questionnaire

will determine future policy.

For example, I received $1 with a survey regarding

my satisfaction with a recent purchase.

The $1 payment was made to make me feel guilty enough

to fill out the questionnaire.

As another example, a city may send out

questionnaires to households and state in a cover letter

that the responses to the questionnaire

will be used to decide pending issues within the city.

Let's consider the Literary Digest poll again.

The Literary Digest mailed out more than 10

million questionnaires and 2.3 million people responded.

The rather low response rate contributed

to the incorrect prediction.

After all, Roosevelt was the incumbent President,

and only those who are unhappy with his administration

were likely to respond.

By the way, in the same election,

the 35-year-old George Gallup predicted

that Roosevelt would win the election

in a survey involving 50,000.

When does nonresponse bias exist?

**3. Non-response can occur because individuals selected**

**for the sample do not wish to respond**

**or the interviewer was unable to contact them.**

1. List two causes of nonresponse bias.

Younger age groups have lower response rates.

People are not available during the time the survey was attempted.

1. List one tool that can be used to control nonresponse bias?

Cash Incentive.

Objective 1, Page 4

 *Answer the following after* *watching the video.*

Under what conditions does response bias exist? RESPONSE BIAS

Response bias exists when the answers on a survey

do not reflect their true feelings of the respondent.

Response bias can occur in a number of ways.

A trained interviewer is essential to obtain

accurate information from a survey.

A skilled interviewer can elicit responses from individuals

and make the interviewee feel comfortable enough

to give truthful responses.

For example, a good interviewer can obtain truthful answers

to questions as sensitive as, have you ever cheated

on your taxes?

Do not be quick to trust surveys conducted

by poorly trained interviewers.

Do not trust survey results if the sponsor has

a vested interest in the results of the survey.

Would you trust the survey conducted

by a car dealer that reports 90% of customers

say they would buy another car from the dealer?

Some survey questions result in responses

that misrepresent facts or are flat out lies.

For example, a survey of recent college graduates

may find that self-reported salaries are inflated.

Also, people may overestimate their abilities.

For example, ask people how many push

ups they can do in one minute.

And then ask them to do the push ups.

How accurate were there?

The way a question is worded can lead

to response bias in a survey, so questions must always

be asked in balance form.

For example, the yes/no question,

INCORRECT

**do you oppose the reduction of the state taxes**

should be written,

CORRECT

**do you favor or oppose**

**the reduction of the state taxes?**

The second question is balanced.

Do you see the difference?

Consider the following report based

on studies from Schuman and Presser

who asked the following two questions—

1. do you think the United States should forbid

public speeches against democracy?

1. Do you think the United States should allow public speeches

against democracy?

**For those respondents presented with question A,**

**21.4% gave yes responses.**

**While for those given question B, 47.8% gave no responses.**

**The conclusion you may arrive that is that most people are**

**not necessarily willing to forbid something but more**

**people are willing not to allow something.**

These results illustrate how wordy a question

can alter a survey's outcome.

Another consideration in wording a question is not to be vague.

The question,

**How much do you study?**

**This question is too vague!**

**Does the researcher mean how much?**

**Do I study for all my classes or just for statistics?**

**Does the researcher mean per day or per week?**

**The question should be written, how many hours**

**do you study statistics each week?**

**Many surveys will rearrange the order of the questions**

**within a questionnaire so the responses are not**

**affected by prior questions.**

**Consider an example from Schuman and Presser**

**in which of the following two questions were asked,**

1. **Do you think the United States should let Communist newspaper**

**reporters from other countries come in here**

**and send back to their papers the news as they see it?**

1. **Do you think a Communist country such as Russia**

**should let American newspaper reporters come in**

**and send back to America the news as they see it?**

***For surveys conducted in 1980 in which the questions appeared***

***A then B, 54.7% of respondents answered yes to A***

***and 63.7% answered yes to B.***

***If the questions were ordered***

***B then A, then 74.6% percent answered yes to A***

***and 81.9% answered yes to B.***

**When Americans are first asked**

**if US reporters should be allowed to report**

**Communist news, they are more likely to agree that Communists**

**should be allowed to report American news.**

**Questions should be rearranged as much as**

**possible to help reduce effects of this type.**

Pollsters will also rearrange words within a question.

For example, the Gallup organization routinely

asked the following question of 1,017 adults age 18 or older.

**The words "approve" and "disapprove"**

**are rotated to remove the effect that**

**may occur by writing the word "approve" first**

**in the question.**

One of the first considerations in designing a question

is determining whether the question

should be open or closed.

An open question allows the responded

to choose his or a response.

**For example, what is the most important problem**

**facing America's youth today?**

**A closed question requires the respondent**

**to choose from a list of predetermined of responses.**

**For example, what is the most important problem**

**facing America's youth today?**

And five answers are given.

**Note:** Response bias can occur through interviewer error, misrepresented answers, wording of questions, ordering of questions or words, type of question, or data-entry error. **A closed question requires the respondent**

**to choose from a list of predetermined of responses.**

**For example, what is the most important problem**

**facing America's youth today?**

And five answers are given.

**CLOSED QUESTIONS**

In closed questions, the possible responses

should be rearranged, because respondents

are likely to choose early choices in a list

rather later choices.

An open question should be phrased

so that the responses are similar.

You don't want a wide variety of responses.

This allows for easy analysis of the responses.

Closed questions limit the number of responded choices,

and therefore, the results are much easier to analyze.

The limited choices, however, do not always

include a respondent's desired choice.

In that case, the respondent will

have to choose a secondary answer or skip the question.

Survey designers recommend conducting pre-test surveys

with open questions and then using the most popular answers

as the choices on closed question surveys.

Another issue to consider the closed question design

is the number of possible responses.

The option "no opinion" should be

omitted, because this option does not

allow for meaningful analysis.

The goal is to limit the number of choices in a closed question

without forcing respondents to choose an option they do not

prefer, which would make the survey have response bias.

**DATA ERROR**

Although not technically a result of response bias,

data entry error will lead to results

that are not representative of the population.

Once data are collected, the results typically

must be entered into a computer which

could result in input errors.

For example, 39 maybe entered as 93.

It is imperative that data be checked for accuracy.

In this text, we present some suggestions

for checking for data error.

**Note:** An open questionallows the respondent to choose his or her response (free response).

**Note:** A closed question requires the respondent to choose from a list of predetermined responses (multiple choice).

Objective 1, Page 7

**Note: Can a Census Have Bias?**

A question on a census form could be misunderstood, thereby leading to response bias in the results. It is often difficult to contact each individual in a population. For example, the U.S. Census Bureau is challenged to count each homeless person in the country, so the census data published by the U.S. A question on a census form could be misunderstood, thereby leading to response bias in the results. It is often difficult to contact each individual in a population. For example, the U.S. Census Bureau is challenged to count each homeless person in the country, so the census data published by the U.S. government likely suffers from nonresponse bias.

Objective 1, Page 8

Define the following terms.

1. Nonsampling Error:

Sampling error:

**Can a Census Have Bias?**

Yes!

The discussion so far has focused on bias in samples, but bias can also occur when conducting a census.

How?

A question on a census form could be misunderstood, thereby leading to response bias in the results.

We also mentioned that it is often difficult to contact each individual in a population. For example, the U.S. Census Bureau faces challenges in counting each homeless person in the country, so the census data published by the U.S. government likely suffers from nonresponse bias.

**Sampling Error versus Nonsampling Error**

Nonresponse bias, response bias, and data-entry errors are types of *nonsampling error*.

However, when a sample is used to learn information about a population, *sampling error* is also likely to occur.

**DEFINITION**

**Nonsampling errors** result from undercoverage, nonresponse bias, response bias, or data-entry error. Such errors could also be present in a census.

**Sampling error** 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.

**In Other Words**

[](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj1_8_eaa86c3e-b4f5-584d-c05f-c0f87a607082)

What do we mean when we say sampling results in "incomplete information"? We mean that the individuals in the sample cannot reveal all the information about the population. Suppose we want to determine the average age of students enrolled in an introductory statistics course. To do this, we find a simple random sample of four students and ask them to write their age on a sheet of paper and turn it in. The average age of these four students is 23.3 years. Assume that no students lied about their age or misunderstood the question and the sampling was done appropriately. If the actual average age of all 30 students in the class (the population) is 22.9 years, then the sampling error is 23.3−22.9=0.4 year. Now suppose the same survey is conducted again, but this time one student lies about his age. The results of that survey will also have nonsampling error.

**Section 1.6 The Design of Experiments**

**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

Introduction, Page 1

 *Watch the video for a review of the language used in observational studies.*

Review the definitions of cross-sectional studies, case-control studies, and cohort studies.

* In observational studies, we cannot make statements of *causality* between the explanatory variable(s) and the response variable.
* The response variable measures the outcome of the study.
* The explanatory variable is the variable whose impact we want to see has on the response variable.

***Objective 1: Describe the Characteristics of an Experiment***

Objective 1, Page 1

 *Define the following terms after watching the video.*

Experiment: An experiment is a controlled study.

**so that's the big difference between it**

**and an observational study.**

It's a controlled study conducted

to determine the effect of varying,

meaning changing, one or more explanatory variables.

And in an experimental design sense,

instead of saying explanatory variables,

Factor: you'll often hear us say factors.

Factors and explanatory variables

1. are synonymous

Treatment: as on a response variable.

Any combination of values and factors

is going to be called a **treatment.**

**So when you have only one explanatory variable,**

**the factors and the treatments are one in the same.**

**But sometimes we might have more than one explanatory variable.**

Let's say we say smoking versus non-smoking is one variable,

and then we might say low protein diet and high protein

diet.

There the treatments would be non-smoker, low protein,

non-smoker, high protein.

Smoker, low protein, and then smoker, high protein.

There's 4 treatments, basically, that

are the combination of factors.

1. Objective 1, Page 1 (continued)

*Define the following terms after watching the video.*

Experimental unit: **The experimental unit** is a person, object,

or some other well defined item upon which

a treatment is applied.

Typically we use experimental unit

when we're not talking about humans.

When we're talking about **humans,** we refer to them as **subjects**,

because it sounds better, I guess.

In a design experiment, you're always

going to have a control group.

Control group: **A control group serves as a baseline treatment**

**that can be used to compare other treatments.**

**Like if I was going to be highly unethical**

**and have this smoking experiment,**

my control group would probably be the nonsmokers,

and then the other group would be hey,

I'm going to make you smoke a pack of cigarettes every day

for the next 30 years and see what happens.

Again, a highly unethical experiment

that hopefully conveys the idea.

Placebo: **A placebo** is another word used in experimental designs.

A placebo is an innocuous medication,

such as a sugar tablet, that looks, tastes, and smells

like the experimental medication.

So maybe I'm trying to do an experiment

to see the impact of alcohol on reaction time.

My placebo there would have to be something

that looks, tastes, and smells like alcohol,

but does not have the effects that alcohol has.

Blinding: **Blinding** refers to nondisclosure of the treatment

an experimental unit is receiving.

So I'm doing an experiment where I'm

trying to see the impact alcohol has on reaction time,

I'm not going to tell you hey, you're getting the real alcohol

and over there, you're getting the placebo.

That would kind of ruin the experiment.

Single-blind: **A single blind**

experiment is one where the experimental unit

does not know the treatment he or she is receiving,

but the researcher or the individual administering

the treatment does.

Those are dicey.

Sometimes it's necessary to do this,

but more times, experiments are done

as a

Double-blind: **double blind experiment**.

Meaning the experimental unit doesn't

know the treatment they're receiving,

nor does the individual administering

the treatment know which treatment the experimental unit

is receiving.

So obviously there needs to be a third party in the background

that does know, but is not involved directly

with the experimental units.

And the reason double blinding is important,

especially in medical studies, is

let's say I am doing an experiment on human subjects

where it's a new medication for the treatment of, say, cancer.

Some type of cancer.

I might, if I know that Bob over here

is getting the new medication, I might, knowingly

or unknowingly, treat Bob different than John

over here who is getting the placebo.

Maybe I'm not a very good poker player

and I show my hands to Bob, and I may also

show my hand to John, and John's like, oh man,

I'm getting the placebo.

I can tell by the way he's acting towards me.

Because there's all kinds of psychosomatic stuff

that oftentimes happens in these medical studies.

And if John knows he's getting a placebo

he might get all down on himself,

and it might be the psychosomatic results that

are impacting the cancer treatment as opposed

to the experimental drug.

Objective 1, Page 2

The use of placebos in designed experiments is a way to form a control group in a designed experiment.

What is the placebo effect? For example, a procedure called **Vertebroplasty** where medical cement is pumped into a spine fracture was tested through a designed experiment. All subjects went through a surgery to repair the spine, but only half received the medical cement. An interesting outcome results from the vertebroplasty experiment. A subject in the placebo group found that the procedure resulted in complete abatement of the back pain even though she did not receive the medical cement! This type of phenomena in an experiment is referred to as the *placebo effect*. A book entitled *Cure* by Jo Marchant explores the **placebo effect**. In the book, she suggests that placebo treatments can have measurable effects. For example, in patients with Parkinson's disease placebos caused an increase of the neurotransmitter dopamine. In a study of 459 migraine sufferers, it was found that the placebo effect accounted for about 60% of the benefit of the drug Maxalt. Of course, the placebo effect will not account for improvements in someone with a tumor or replace insulin with someone with diabetes. However, the Maxalt study suggests that remedies for pain, nausea, or depression rely extensively on the placebo effect.

Objective 1, Page 3

Recall confounding in a study occurs when the effects of two or more explanatory variables are not separated. In designed experiments, confounding may occur as a result of a confounding variable, which is an explanatory variable that was considered in a study whose effect cannot be distinguished from a second explanatory variable in the study. The example we used to illustrate the concept of confounding was Professor Egner's study on the effect of online homework versus paper/pencil homework on final exam scores. Professor Egner taught her morning class using online homework and her afternoon class using paper/pencil homework. If the final exam scores for the morning class were higher than the afternoon class, we could not tell whether the higher exam scores are a result of the homework system, or the time the class is offered. Therefore, the explanatory variable homework system is confounded with the explanatory variable time of day.

Well-designed experiments will account for the potential of confounding in a study.

Objective 1, Page 6

**Example 1 *The Characteristics of an Experiment***

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 whether there was an occurrence of any major cardiovascular event or not.

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.

What does it mean for the experiment to be placebo-controlled? 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.

What does it mean for the experiment to be placebo-controlled?

What does it mean for the experiment to be double-blind? 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.

What does it mean for the experiment to be double-blind?

What is the population for which this study applies? What is the sample? 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.

Objective 1, Page 6 (continued)

* + 1. What are the treatments?
    2. What is the response variable? Is it qualitative or quantitative?

***Objective 2: Explain the Steps in Designing an Experiment***

Objective 2, Page 1

**Steps in Conducting a Designed Experiment**

Fill in each step.

***Step 1***: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The statement of the problem should be as explicit as possible and should provide the experimenter with direction. The statement must also identify the response variable and the population to be studied. Often, the statement is referred to as the *claim.*

***Step 2***: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The factors are usually identified by an expert in the field of study. In identifying the factors, ask, “What things affect the value of the response variable?” After the factors are identified, determine which factors to fix at some predetermined level, which to manipulate, and which to leave uncontrolled.

***Step 3***: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

As a general rule, choose as many experimental units as time and money allow. Techniques exist for determining sample size, provided certain information is available.

Objective 2, Page 1 (continued)

***Step 4****:* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Factors can be dealt with in two ways - control or randomize.

Controlmeans to either set the factor at one value throughout the experiment or set the level of the factor at various levels).

Randomizemeans to randomly assign the experimental units to various treatment groups.

***Step 5****:* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Replicationoccurs when each treatment is applied to more than one experimental unit.

***Step 6****:* \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Inferential statisticsis a process in which generalizations about a population are made on the basis of results obtained from a sample.

Objective 2, Page 2

List the six steps for the Lipitor study in Example 1 (Objective 1, Page 6)

**Step 1:** *Identify the Problem to be Solved*

**Step 2*:*** *Determine the Factors That Affect the Response Variable*

**Step 3:** *Determine the Number of Experimental Units*

**Step 4:** *Determine the Level of Each Factor*

**Step 5*:*** *Conduct the Experiment*

**Step 6:** *Test the Claim*

***Objective 3: Explain the Completely Randomized Design***

Objective 3, Page 1

1. What is a completely randomized design?

Objective 3, Page 2

**Example 2 *A Completely Randomized Design***

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

Objective 3, Page 3

Sketch the experimental design from Example 2 (Objective 3, Page 2).

1. Explain why this experimental design is a completely randomized design.

***Objective 4: Explain the Matched-Pairs Design***

Objective 4, Page 1

1. What is a matched-pairs design?

The pairs are selected so that they are related in some way.

There are only two levels of treatment in a matched-pairs design.

Objective 4, Page 2

**Example 3 *A Matched-Pairs Design***

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

What is the population for which this study applies? What is the sample?

Solution

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

What are the treatments?

Solution

The treatments are 10 mg of Lipitor or a placebo daily

What is the response variable? Is it qualitative or quantitative?

Solution

The response variable is whether the subject had any major cardiovascular event, such as a stroke, or not.

It is a qualitative variable with two possible outcomes—cardiovascular event or not

**Steps in Designing an Experiment**

1. *Identify the Problem to Be Solved.* The statement of the problem should be as explicit as possible and should provide the experimenter with direction. The statement must also identify the response variable and the population to be studied. Often, the statement is referred to as the *claim.*
2. *Determine the Factors That Affect the Response Variable*. The factors are usually identified by an expert in the field of study. In identifying the factors, ask, “What things affect the value of the response variable?” After the factors are identified, determine which factors to fix at some predetermined level, which to manipulate, and which to leave uncontrolled
3. *Determine the Number of Experimental Units*. As a general rule, choose as many experimental units as time and money allow. Techniques exist for determining sample size, provided certain information is available.

*Determine the Level of Each Factor.* There are two ways to deal with the factors, control or randomize.

1. **Control:** There are two ways to control the factors.
   1. Set the level of a factor at one value throughout the experiment (if you are not interested in its effect on the response variable).
   2. Set the level of a factor at various levels (if you are interested in its effect on the response variable). The combinations of the levels of all varied factors constitute the treatments in the experiment.

**Randomize:** Randomly assign the experimental units to treatment groups.  Because it is difficult, if not impossible, to identify all factors in an experiment, randomly assigning experimental units to treatment groups reduces the effect of variation attributable to factors (explanatory variables) not controlled. That is, randomly assigning experimental units to treatment groups tends to "even out" any uncontrolled factors.

*Determine the Level of Each Factor.* There are two ways to deal with the factors, control or randomize.

1. **Control:** There are two ways to control the factors.
   1. Set the level of a factor at one value throughout the experiment (if you are not interested in its effect on the response variable).
   2. Set the level of a factor at various levels (if you are interested in its effect on the response variable). The combinations of the levels of all varied factors constitute the treatments in the experiment.
2. **Randomize:** Randomly assign the experimental units to treatment groups.  Because it is difficult, if not impossible, to identify all factors in an experiment, randomly assigning experimental units to treatment groups reduces the effect of variation attributable to factors (explanatory variables) not controlled. That is, randomly assigning experimental units to treatment groups tends to "even out" any uncontrolled factors.
3. ***Conduct the Experiment.***
4. Replication occurs when each treatment is applied to more than one experimental unit. Using more than one experimental unit for each treatment ensures the effect of a treatment is not due to some characteristic of a single experimental unit. It is a good idea to assign an equal number of experimental units to each treatment.
5. Collect and process the data. Measure the value of the response variable for each replication. Then organize the results. The idea is that the value of the response variable for each treatment group is the same before the experiment because of randomization. Then any difference in the value of the response variable among the different treatment groups is a result of differences in the level of the treatment.
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
   1. To help understand the steps in designing an experiment, let's review [Example 1](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_2_a0c604ad-5b50-fa67-9c74-c382df6b6d1d).

*Identify the Problem to Be Solved*. The problem to be solved is to determine whether 10 mg of Lipitor daily reduces the likelihood of having a major cardiovascular event in 40- to 75-year old subjects with type 2 diabetes.

* 1. *Determine the Factors That Affect the Response Variable*. Some factors that may affect whether one has a cardiovascular event are diet, exercise, family history, and level of cholesterol.

*3.Determine the Number of Experimental Units.* There were 2838 subjects in the study.

*4.Determine the Level of Each Factor.* The factor of interest is the drug, which was set at two levels: placebo and 10 mg of Lipitor. Although not stated, the researchers likely fixed the diet of the subjects and fixed an exercise regimen. Family history cannot be controlled, so the random assignment of the subjects to two groups will average out a bad family history of heart disease. For example, we would not expect all subjects with a poor history of heart health to end up in the placebo (control) group.

*5. Conduct the Experiment.* The subjects were randomly assigned to either the placebo or Lipitor group. There were 2838 replications of the experiment.

*6. Test the Claim*. The inferential statistics suggest that the Lipitor group had a lower rate of major cardiovascular events

e now concentrate on the simplest type of experiment.

**DEFINITION**

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

The study from Example 1 is a completely randomized design because each experimental unit (the 2838 subjects) was randomly assigned to either the placebo group or Lipitor group

Problem

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

**Video Solution**

|  |
| --- |
|  |

Approach

Follow the [steps](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj3_2_fcab5ca6-ac0a-bdc3-ea57-3afa4f9b30d5) for designing an experiment.

Solution

**Step 1. Identify the Problem to Be Solved** 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.

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

**Determine the Number of Experimental Units** In this experiment, we will plant 60 soybean plants (experimental units).

**Determine the Level of Each Factor** List the factors and their levels.

* **Fertilizer.** This factor is the explanatory variable of interest. So, it 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.
* **Precipitation.**The amount of rainfall cannot be controlled, but the amount of watering done can be controlled. Each plant will receive the same amount of precipitation.
* **Sunlight.** This uncontrollable factor will be roughly the same for each plant.
* **Method of tilling.**Control this factor using the round-up ready method of tilling for each plant.
* **Type of soil.**Certain aspects of the soil, such as level of acidity, can be controlled. In addition, each plant will be planted within a 1-acre area, so it is reasonable to assume that the soil conditions for each plant are equivalent.
* **Plant.** There may be variation (in terms of ability to generate yield) from plant to plant. To account for this, we randomly assign the plants to a treatment.
* **Temperature.** This factor is uncontrollable, but will be the same for each plant.

**Conduct the Experiment**

1. Assign each plant to a treatment group. First, number the plants from 1 to 60 and randomly generate 20 numbers. The plants corresponding to these numbers get treatment A. Next number the remaining plants 1 to 40 and randomly generate 20 numbers. The plants corresponding to these numbers get treatment B. The remaining plants get treatment C. Now till the soil, plant the soybean plants, and fertilize according to the schedule prescribed.
2. At the end of the growing season, determine the crop yield for each plant.

**Test the Claim.**Determine any differences in yield among the three treatment groups

gure 2 illustrates the experimental design from Example 2 on the previous screen.

|  |  |
| --- | --- |
| An experimental design. |  |
| Figure 2 | |

Example 2 is a completely randomized design because the experimental units (the plants) were randomly assigned to the treatments. It is the most popular experimental design because of its simplicity, but it is not always the best.

OBJECTIVE 4 Explain the Matched-Pairs Design

Another type of experimental design is called a *matched-pairs design*.

**DEFINITION**

A **matched-pairs design** is an experimental design in which the experimental units are paired up. The pairs are selected so that they are related in some way (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.

In matched-pairs design, one matched individual will receive one treatment and the other receives a different treatment. The matched pair is randomly assigned to the treatment using a coin flip or a random-number generator. We then look at the difference in the results of each matched pair. One common type of matched-pairs design is to measure a response variable on an experimental unit before and after a treatment is applied. In this case, the individual is matched against itself. These experiments are sometimes called before–after or pretest–posttest experiments.

A Matched-Pairs Design

Problem

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

**Video Solution**

|  |
| --- |
|  |

Approach

Use a matched-pairs design by matching students according to IQ and gender (just in case gender plays a role in learning with music).

Solution

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

For each pair of students, 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 and asked to study Section 1.1. After 2 hours, the students will enter a testing center and take a short quiz on the material in the section. Compute the difference in the scores of each matched pair. Any differences in scores will be attributed to the treatment. Figure 3 illustrates the design.

|  |  |
| --- | --- |
| [Alt-Text](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx)A matched-pairs design.  Figure 3 |  |

The experimental design is based on how the subjects are grouped. Different types of experimental designs and studies are explained below.

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

A​ matched-pairs design is an experimental design in which the experimental units are paired up. The pairs are selected so that they are related in some way​ (that is, the same person before and after a​ treatment, twins, husband and​ wife, same geographical​ location, and so​ on).

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.

​

Case-control studies are retrospective​ studies, 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 a certain characteristic may be matched with those who do not.

An experiment is a controlled study conducted to determine the effect 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.