Chapter 1

Statistical Basics

You are exposed to statistics regularly. If you are a sports fan, then you have the statistics for your favorite player. If you are interested in politics, then you look at the polls to see how people feel about certain issues or candidates. If you are an environmentalist, then you research arsenic levels in the water of a town or analyze the global temperatures. If you are in the business profession, then you may track the monthly sales of a store or use quality control processes to monitor the number of defective parts manufactured. If you are in the health profession, then you may look at how successful a procedure is or the percentage of people infected with a disease. There are many other examples from other areas. To understand how to collect data and analyze it, you need to understand what the field of statistics is and the basic definitions.

1.1 What is Statistics?

Statistics is the study of how to collect, organize, analyze, and interpret data collected from a group.

There are two branches of statistics. One is called descriptive statistics, which is where you collect and organize data. The other is called inferential statistics, which is where you analyze and interpret data. First you need to look at descriptive statistics since you will use the descriptive statistics when making inferences.

To understand how to create descriptive statistics and then conduct inferences, there are a few definitions that you need to look at. Note, many of the words that are defined have common definitions that are used in non-statistical terminology. In statistics, some have slightly different definitions. It is important that you notice the difference and utilize the statistical definitions.

The first thing to decide in a statistical study is whom you want to measure and what you want to measure. You always want to make sure that you can answer the question of whom you measured and what you measured. The who is known as the unit of observation and the what is the variable(s).

Unit of observation – a person or object that you are interested in finding out information about.

Variable – the measurement or observation of the unit of observation

Having the unit of observation and the variables is part of picture of a **data set** or **data frame**. To make a data set or data frame into what is called tidy data, it should be organized in a way that each row of the data frame is a unit of observation, and the variables should be well defined and are easily identified. An example of a data frame that is tidy data is:

```
Sugar <- read.csv("https://krkozak.github.io/MAT160/sugar.csv")
head(Sugar) # Displays the variables and the frist few lines of units of observations.
```

```
##                        name chidren mfr type calories protein fat sodium
## 1                 100%_Bran       N   N    C       70       4   1    130
## 2          100%_Natural_Bran       N   Q    C      120       3   5     15
## 3                   All-Bran       N   K    C       70       4   1    260
## 4 All-Bran_with_Extra_Fiber       N   K    C       50       4   0    140
## 5             Almond_Delight       N   R    C      110       2   2    200
## 6   Apple_Cinnamon_Cheerios       Y   G    C      110       2   2    180
##   fiber carbo sugars potass vitamins shelf weight cups   rating
## 1  10.0   5.0      6    280       25     3      1 0.33 68.40297
## 2   2.0   8.0      8    135        0     3      1 1.00 33.98368
## 3   9.0   7.0      5    320       25     3      1 0.33 59.42551
## 4  14.0   8.0      0    330       25     3      1 0.50 93.70491
## 5   1.0  14.0      8     -1       25     3      1 0.75 34.38484
## 6   1.5  10.5     10     70       25     1      1 0.75 29.50954
```

Collecting multiple variables from one unit of observation makes sense. If you wanted to figure out the diameter of breast height of Ponderosa Pine trees in the Coconino National Forest, you need to physically measure a bunch of trees. While you are measuring the diameter, you might also want to measure the height of the tree, if the tree has a bark beetle infestation, the estimated age of the tree, the color of the bark, and how many branches it has. You may only want to estimate the average diameter at breast height, but now you have the ability to estimate other quantities too. No sense walking all over the forest and only measure one thing.

A large data frame is one that has at least 5 variable and at least 1000 units of observations.if a data frame only has 3 variables and 500 rows, that doesn't make it not usable. The 1000 units of observation and 5 variables is just a guideline to work with.

If you put the unit of observation and the variable into one statement, then you obtain a population.

Population – set of all values of the variable for the entire group of individuals.

Notice, the population answers who you want to measure and what you want to measure. Make sure that your population always answers both of these questions. If it doesn't, then you haven't given someone who is reading your study the entire picture. As an example, if you just say that you are going to collect data from the senators in the U.S. Congress, you haven't told your reader want you are going to collect. Do you want to know their income, their highest degree earned, their voting record, their age, their political party, their gender, their marital status, or how they feel about a particular issue? Without telling what you want to measure, your reader has no idea what your study is actually about.

Sometimes the population is very easy to collect. Such as if you are interested in finding the average age of all of the current senators in the U.S. Congress, there are only 100 senators. This wouldn't be hard to find. However, if instead you were interested in knowing the average age that a senator in the U.S. Congress first took office for all senators that ever served in the U.S. Congress, then this would be a bit more work. It is still doable, but it would take a bit of time to collect. But what if you are interested in finding the average diameter of breast height of all of the Ponderosa Pine trees in the Coconino National Forest? This would be impossible to actually collect. What do you do in these cases? Instead of collecting the entire population, you take a smaller group of the population, kind of a snap shot of the population. This smaller group is called a sample.

Sample – a subset from the population. It looks just like the population, but contains less data.

In today of big data, there is some confusion between really large data frames and populations. The population is a theoretical concept and even if you have a very large data frame, that doesn't mean you have the population. Most populations are not actually able to be collected. They are considered an ideal that you are trying to make decisions about.

How you collect your sample can determine how accurate the results of your study are. There are many ways to collect samples. Some of them create better samples than others. No sampling method is perfect, but some are better than others. Sampling techniques will be discussed later. For now, realize that every time you take a sample you will find different data values. The sample is a snapshot of the population, and there is more information than is in the picture. The idea is to try to collect a sample that gives you an accurate picture, but you will never know for sure if your picture is the correct picture. Unlike previous mathematics classes where there was always one right answer, in statistics there can be many answers, and you don't know which are right.

Once you have your data frame, either from a population or a sample, you need to know how you want to summarize the data. As an example, suppose you are interested in finding the proportion of people who like a candidate, the average height a plant grows to using a new fertilizer, or the variability of the test scores. Understanding how you want to summarize the data helps to determine the type of data you want to collect. Since the population is what we are interested in, then you want to calculate a number from the population. This is known as a parameter. As mentioned already, you can't really collect the entire population. Even though this is the number you are interested in, you can't really calculate it. Instead you use a number calculated from the sample, called a statistic, to estimate the parameter. Since no sample is exactly the same, the statistic values are going to be different from sample to sample. They estimate the value of the parameter, but again, you do not know for sure if your answer is correct.

Parameter – a number calculated from the population. Usually denoted with a Greek letter. This number is a fixed, unknown number that you want to find.

Statistic – a number calculated from the sample. Usually denoted with letters from the Latin alphabet, though sometimes there is a Greek letter with a $\widehat{\ }$ (called a hat) above it. Since you can find samples, it is readily known, though it changes depending on the sample taken. It is used to estimate the parameter value.

One last concept to mention is that there are two different types of variables – qualitative (categorical) and quantitative (numerical). Each type of variable has different parameters and statistics that you find. It is important to know the difference between them.

Qualitative or categorical variable – answer is a word or name that describes a quality of the individual.

Quantitative or numerical variable – answer is a number, something that can be counted or measured from the individual.

1.1.1 Example: Stating Definitions for Qualitative Variable

In 2010, the Pew Research Center questioned 1500 adults in the U.S. to estimate the proportion of the population favoring marijuana use for medical purposes. It was found that 73% are in favor of using > marijuana for medical purposes. State the unit of observation, variable, population, and sample.

Solution:

Unit of observation – a U.S. adult

Variable – the response to the question "should marijuana be used for medical purposes?" This is qualitative data since you are recording a > person's response – yes or no.

Population – set of all responses of adults in the U.S.

Sample – set of 1500 responses of U.S. adults who are questioned.

Parameter – proportion of those who favor marijuana for medical purposes calculated from population

Statistic– proportion of those who favor marijuana for medical > purposes calculated from sample

1.1.2 Example: Stating Definitions for Qualitative Variable

A parking control officer records the manufacturer of every 5^{th} car in the college parking lot in order to guess the most common manufacturer. State the unit of observation, variable, population, and sample.

Solution:

Unit of observation – a car in the college parking lot

Variable – the name of the manufacturer. This is qualitative data since you are recording a car type.

Population – set of all names of the manufacturer of cars in the college parking lot.

Sample – set of recorded names of the manufacturer of the cars in college parking lot

Parameter – proportion of each car type calculated from population

Statistic – proportion of each car type calculated from sample

1.1.3 Example: Stating Definitions for Quantitative Variable

A biologist wants to estimate the average height of a plant that is > given a new plant food. She gives 10 plants the new plant food. State the unit of observation, variable, population, and sample.

Solution:

Unit of observation – a plant given the new plant food

Variable – the height of the plant (Note: it is not the average > height since you cannot measure an average – it is calculated from data.) This is quantitative data since you will have a number.

Population – set of all the heights of plants when the new plant food is used

Sample – set of 10 heights of plants when the new plant food is used

Parameter – average height of all plants when the new plant food is used

Statistic – average height of 10 plants when the new plant food is used

Note: in example #1.1.3, you most likely will be comparing the new plant food to either an old plant food. So you would have more units of observations, but for plants given the old plant food.You may also want to have measurements on particular days after you give the plant food. In your data frame you would need to have many variables besides just the height of the plant. Examples of variables would be plant_number, fertilizer (yes or no), height on day 1, height on day 30, height on day 60, and so forth. One other comment, you variable names should make sense to your reader, and be one word for ease in analyzing by a computer program.

1.1.4 Example: Stating Definitions for Quantitative Variable

A doctor wants to see if a new treatment for cancer extends the life expectancy of a patient versus the old treatment. She gives one group of 25 cancer patients the new treatment and another group of 25 the old treatment. She then measures the life expectancy of each of the patients. State the individuals, variables, populations, and samples.

Solution:

In this example there are two individuals, two variables, two populations, and two samples.

Individual 1: cancer patient given new treatment

Individual 2: cancer patient given old treatment

Variable 1: life expectancy when given new treatment. This is quantitative data since you will have a number.

Variable 2: life expectancy when given old treatment. This is quantitative data since you will have a number.

Population 1: set of all life expectancies of cancer patients given new treatment

Population 2: set of all life expectancies of cancer patients given old treatment

Sample 1: set of 25 life expectancies of cancer patients given new treatment

Sample 2: set of 25 life expectancies of cancer patients given old treatment

Parameter 1 – average life expectancy of all cancer patients given new treatment

Parameter 2 – average life expectancy of all cancer patients given old treatment

Statistic 1 – average life expectancy of 25 cancer patients given new treatment

Statistic 2 – average life expectancy of 25 cancer patients given old treatment

There are different types of quantitative variables, called discrete or continuous. The difference is in how many values can the data have. If you can actually count the number of data values (even if you are counting to infinity), then the variable is called discrete. If it is not possible to count the number of data values, then the variable is called continuous.

Discrete data can only take on particular values like integers. Discrete data are usually things you count.

Continuous data can take on any value. Continuous data are usually things you measure.

1.1.5 Example: Discrete or Continuous

Classify the quantitative variable as discrete or continuous.

a.) The weight of a cat.

Solution:

This is continuous since it is something you measure.

b.) The number of fleas on a cat.

Solution:

This is discrete since it is something you count.

c.) The size of a shoe.

Solution:

This is discrete since you can only be certain values, such as 7, 7.5, 8, 8.5, 9. You can't buy a 9.73 shoe.

There are also are four measurement scales for different types of data with each building on the ones below it. They are:

Measurement Scales:

Nominal – data is just a name or category. There is no order to any data and since there are no numbers, you cannot do any arithmetic on this level of data. Examples of this are gender, car name, ethnicity, and race.

Ordinal – data that is nominal, but you can now put the data in order, since one value is more or less than another value. You cannot do arithmetic on this data, but you can now put data values in order. Examples of this are grades (A, B, C, D, F), place value in a race (1st, 2nd, 3rd), and size of a drink (small, medium, large).

Interval – data that is ordinal, but you can now subtract one value from another and that subtraction makes sense. You can do arithmetic on this data, but only addition and subtraction. Examples of this are temperature and time on a clock.

Ratio – data that is interval, but you can now divide one value by another and that ratio makes sense. You can now do all arithmetic on this data. Examples of this are height, weight, distance, and time.

Nominal and ordinal data come from qualitative variables. Interval and ratio data come from quantitative variables.

Most people have a hard time deciding if the data are nominal, ordinal, interval, or ratio. First, if the variable is qualitative (words instead of numbers) then it is either nominal or ordinal. Now ask yourself if you can put the data in a particular order. If you can it is ordinal. Otherwise, it is nominal. If the variable is quantitative (numbers), then it is either interval or ratio. For ratio data, a value of 0 means there is no measurement. This is known as the absolute zero. If there is an absolute zero in the data, then it means it is ratio. If there is no absolute zero, then the data are interval. An example of an absolute zero is if you have \$0 in your bank account, then you are without money. The amount of money in your bank account is ratio data. *Word of caution*: sometimes ordinal data is displayed using numbers, such as 5 being strongly agree, and 1 being strongly disagree. These numbers are not really numbers. Instead they are used to assign numerical values to ordinal data. In reality you should not perform any computations on this data, though many people do. If there are numbers, make sure the numbers are inherent numbers, and not numbers that were assigned.

1.1.6 Example: Measurement Scale

State which measurement scale each is.

a.) Time of first class

Solution:

This is interval since it is a number, but 0 o'clock means midnight and not the absence of time.

b.) Hair color

Solution:

This is nominal since it is not a number, and there is no specific order for hair color.

c.) Length of time to take a test

Solution:

This is ratio since it is a number, and if you take 0 minutes to take a test, it means you didn't take any time to complete it.

d.) Age groupings (baby, toddler, adolescent, teenager, adult, elderly)

Solution:

This is ordinal since it is not a number, but you could put the data in order from youngest to oldest or the other way around.

1.1.7 Homework

1. Suppose you want to know how Arizona workers age 16 or older travel to work. To estimate the percentage of people who use the different modes of travel, you take a sample containing 500 Arizona workers age 16 or older. State the individual, variable, population, sample, parameter, and statistic.

2. You wish to estimate the mean cholesterol levels of patients two days after they had a heart attack. To estimate the mean you collect data from 28 heart patients. State the individual, variable, population, sample, parameter, and statistic.

3. Print-O-Matic would like to estimate their mean salary of all employees. To accomplish this they collect the salary of 19 employees. State the individual, variable, population, sample, parameter, and statistic.

4. To estimate the percentage of households in Connecticut which use fuel oil as a heating source, a researcher collects information from 1000 Connecticut households about what fuel is their heating source. State the individual, variable, population, sample, parameter, and statistic.

5. The U.S. Census Bureau needs to estimate the median income of males in the U.S., they collect incomes from 2500 males. State the individual, variable, population, sample, parameter, and statistic.

6. The U.S. Census Bureau needs to estimate the median income of females in the U.S., they collect incomes from 3500 females. State the individual, variable, population, sample, parameter, and statistic.

7. Eyeglassmatic manufactures eyeglasses and they would like to know the percentage of each defect type made. They review 25,891 defects and classify each defect that is made. State the individual, variable, population, sample, parameter, and statistic.

8. The World Health Organization wishes to estimate the mean density of people per square kilometer, they collect data on 56 countries. State the individual, variable, population, sample, parameter, and statistic

9. State the measurement scale for each.

a. Cholesterol level
b. Defect type
c. Time of first class
d. Opinion on a 5 point scale, with 5 being strongly agree and 1 being strongly disagree

10. State the measurement scale for each.

a. Temperature in degrees Celsius
b. Ice cream flavors available
c. Pain levels on a scale from 1 to 10, 10 being the worst pain ever
d. Salary of employees

1.2 Sampling Methods

> As stated before, if you want to know something about a population, it is often impossible or impractical to examine the whole population. It might be too expensive in terms of time or money. It might be impractical – you can't test all batteries for their length of lifetime because there wouldn't be any batteries left to sell. You need to look at a sample. Hopefully the sample behaves the same as the population.

When you choose a sample you want it to be as similar to the population as possible. If you want to test a new painkiller for adults you would want the sample to include people who are fat, skinny, old, young, healthy, not healthy, male, female, etc.

There are many ways to collect a sample. None are perfect, and you are not guaranteed to collect a representative sample. That is unfortunately the limitations of sampling. However, there are several techniques that can result in samples that give you a semi-accurate picture of the population. Just remember to be aware that the sample may not be representative. As an example, you can take a random sample of a group of people that are equally males and females, yet by chance everyone you choose is female. If this happens, it may be a good idea to collect a new sample if you have the time and money. > There are many sampling techniques, though only four will be presented here. The simplest, and the type that is Desired for is a **simple random sample**. This is where you pick the sample such that every sample has the same chance of being chosen. This type of sample is actually hard to collect, since it is sometimes difficult to obtain a complete list of all individuals. There are many cases where you cannot conduct a truly random sample. However, you can get as close as you can. Now suppose you are interested in what type of music people like. It might not make sense to try to find an answer for everyone in the U.S. You probably don't like the same music as your parents. The answers vary so much you probably couldn't find an answer for everyone all at once. It might make sense to look at people in different age groups, or people of different ethnicities. This is called a **stratified sample**. The issue with this sample type is that sometimes people subdivide the population too much. It is best to just have one stratification. Also, a stratified sample has similar problems that a simple random sample has. If your population has some order in it, then you could do a **systematic sample**. This is popular in manufacturing. The problem is that it is possible to miss a manufacturing mistake because of how this sample is taken. If you are collecting polling data based on location, then a **cluster sample** that divides the population based on geographical means would be the easiest sample to conduct. The problem is that if you are looking for opinions of people, and people who live in the same region may have similar opinions. As you can see each of the sampling techniques have pluses and minuses. Include convenience [NOTE IN DRAFT: This sentence is incomplete.]

A **simple random sample (SRS)** of size n is a sample that is selected from a population in a way that ensures that every different possible sample of size n has the same chance of being selected. Also, every individual associated with the population has the same chance of being selected. Ways to select a simple random sample:

– Put all names in a hat and draw a certain number of names out. – Assign each individual a number and use a random number table or a calculator or computer to randomly select the individuals that will be measured.

1.2.1 Example: Choosing a Simple Random Sample

Describe how to take a simple random sample from a classroom.

Solution:

Give each student in the class a number. Using a random number generator you could then pick the number of students you want to pick.

1.2.2 Example: How Not to Choose a Simple Random Sample

You want to choose 5 students out of a class of 20. Give some examples of samples that are not simple random samples:

Solution:

Choose 5 students from the front row. The people in the last row have no chance of being selected. Choose the 5 shortest students. The tallest students have no chance of being selected. Ask your friend to pick numbers that have been assinged to each student. Your friend may prefer certain numbers and picks those. This is not knownn by your friend, but this happens.

1.2.3 Example: How to Choose a Simple Random Sample using R

You want to take a simple random sample of size 1000 from a dataframe known as NHANES.

Solution:

```
library("NHANES") # turns on the package NHANES in R
sample_NHANES<-sample(NHANES, 1000) #creates a random sample and saves it as Sample_NHANES
sample_NHANES #displays the sample just created
```

```
## # A tibble: 1,000 x 77
##       ID SurveyYr Gender   Age AgeDecade AgeMonths Race1 Race3 Education
##    <int> <fct>    <fct>  <int> <fct>         <int> <fct> <fct> <fct>
##  1 66632 2011_12  male      58 " 50-59"         NA White White 9 - 11th G~
##  2 64806 2011_12  male      80 <NA>             NA White White Some Colle~
##  3 56753 2009_10  male      61 " 60-69"        735 White <NA>  College Gr~
##  4 55788 2009_10  female    26 " 20-29"        313 White <NA>  College Gr~
##  5 64710 2011_12  male       3 " 0-9"           NA Other Other <NA>
##  6 67578 2011_12  female    10 " 10-19"         NA White White <NA>
##  7 66243 2011_12  female    39 " 30-39"         NA Black Black Some Colle~
##  8 56240 2009_10  male      53 " 50-59"        639 White <NA>  College Gr~
```

```
##  9 58611 2009_10  male      48 " 40-49"        587 White <NA>  College Gr~
## 10 57571 2009_10  female    70 " 70+"          844 Other <NA>  High School
## # ... with 990 more rows, and 68 more variables: MaritalStatus <fct>,
## #   HHIncome <fct>, HHIncomeMid <int>, Poverty <dbl>, HomeRooms <int>,
## #   HomeOwn <fct>, Work <fct>, Weight <dbl>, Length <dbl>, HeadCirc <dbl>,
## #   Height <dbl>, BMI <dbl>, BMICatUnder20yrs <fct>, BMI_WHO <fct>,
## #   Pulse <int>, BPSysAve <int>, BPDiaAve <int>, BPSys1 <int>,
## #   BPDia1 <int>, BPSys2 <int>, BPDia2 <int>, BPSys3 <int>, BPDia3 <int>,
## #   Testosterone <dbl>, DirectChol <dbl>, TotChol <dbl>, UrineVol1 <int>,
## #   UrineFlow1 <dbl>, UrineVol2 <int>, UrineFlow2 <dbl>, Diabetes <fct>,
## #   DiabetesAge <int>, HealthGen <fct>, DaysPhysHlthBad <int>,
## #   DaysMentHlthBad <int>, LittleInterest <fct>, Depressed <fct>,
## #   nPregnancies <int>, nBabies <int>, Age1stBaby <int>,
## #   SleepHrsNight <int>, SleepTrouble <fct>, PhysActive <fct>,
## #   PhysActiveDays <int>, TVHrsDay <fct>, CompHrsDay <fct>,
## #   TVHrsDayChild <int>, CompHrsDayChild <int>, Alcohol12PlusYr <fct>,
## #   AlcoholDay <int>, AlcoholYear <int>, SmokeNow <fct>, Smoke100 <fct>,
## #   Smoke100n <fct>, SmokeAge <int>, Marijuana <fct>, AgeFirstMarij <int>,
## #   RegularMarij <fct>, AgeRegMarij <int>, HardDrugs <fct>, SexEver <fct>,
## #   SexAge <int>, SexNumPartnLife <int>, SexNumPartYear <int>,
## #   SameSex <fct>, SexOrientation <fct>, PregnantNow <fct>, orig.id <chr>
```

> **Stratified sampling** is where you break the population into groups called strata, then take a simple random sample from each strata.

For example:

–If you want to look at musical preference, you could divide the individuals into age groups and then conduct simple random samples inside each group.

–If you want to calculate the average price of textbooks, you could divide the individuals into groups by major and then conduct simple random samples inside each group.

> **Systematic sampling** is where you randomly choose a starting place then select every kth individual to measure.

For example:

–You select every 5^{th} item on an assembly line –You select every 10^{th} name on the list –You select every 3^{rd} customer that comes into the store.

> **Cluster sampling** is where you break the population into groups called clusters. Randomly pick some clusters then poll all individuals in those clusters.

For example:

–A large city wants to poll all businesses in the city. They divide the city into sections (clusters), maybe a square block for each section, and use a random number generator to pick some of the clusters. Then they poll all businesses in each chosen cluster. –You want to measure whether a tree in the forest is infected with bark beetles. Instead of having to walk all over the forest, you divide the forest up into sectors, and then randomly pick the sectors that you will travel to. Then record whether a tree is infected or not for every tree in that sector.

Many people confuse stratified sampling and cluster sampling. In stratified sampling you use all the groups and some of the members in each group. Cluster sampling is the other way around. It uses some of the groups and all the members in each group.

The four sampling techniques that were presented all have advantages and disadvantages. There is another sampling technique that is sometimes utilized because either the researcher doesn't know better, or it is easier to do. This sampling technique is known as a convenience sample. This sample will not result in a representative sample, and should be avoided.

> **Convenience sample** is one where the researcher picks individuals to be included that are easy for the researcher to collect.

An example of a convenience sample is if you want to know the opinion of people about the criminal justice system, and you stand on a street corner near the county court house, and questioning the first 10 people who walk by. The people who walk by the county court house are most likely involved in some fashion with the criminal justice system, and their opinion would not represent the opinions of all individuals.

On a rare occasion, you do want to collect the entire population. In which case you conduct a census.

> A **census** is when every unit of observation is measured.

1.2.4 Example: Sampling type

Banner Health is a several state nonprofit chain of hospitals. Management wants to assess the incident of complications after surgery. They wish to use a sample of surgery patients. Several sampling techniques are described below. Categorize each technique as simple random sample, stratified sample, systematic sample, cluster sample, or convenience sampling.

a. Obtain a list of patients who had surgery at all Banner Health facilities. Divide the patients according to type of surgery. Draw simple random samples from each group.

Solution

This is a stratified sample since the patients where separated into different stratum and then random samples were taken from each strata. The problem with this is that some types of surgeries may have more chances for complications than others. Of course, the stratified sample would show you this.

b. Obtain a list of patients who had surgery at all Banner Health facilities. Number these patients, and then use a random number table to obtain the sample.

Solution

This is a random sample since each patient has the same chance of being chosen. The problem with this one is that it will take a while to collect the data.

c. Randomly select some Banner Health facilities from each of the seven states, and then include all the patients on the surgery lists of the states.

Solution

This is a cluster sample since all patients are questioned in each of the selected hospitals. The problem with this is that you could have by chance selected hospitals that have no complications.

d. At the beginning of the year, instruct each Banner Health facility to record any complications from every 100^{th} surgery.

Solution

This is a systematic sample since they selected every 100^{th} surgery. The problem with this is that if every 90^{th} surgery has complications, you wouldn't see this come up in the data.

e. Instruct each Banner Health facilities to record any complications from 20 surgeries this week and send in the results.

Solution

This is a convenience sample since they left it up to the facility how to do it. The problem with convenience samples is that the person collecting the data will probably collect data from surgeries that had no complications.

1.2.5 Homework

1. Researchers want to collect cholesterol levels of U.S. patients who had a heart attack two days prior. The following are different sampling techniques that the researcher could use. Classify each as simple random sample, stratified sample, systematic sample, cluster sample, or convenience sample.

a. The researchers randomly select 5 hospitals in the U.S. then measure the cholesterol levels of all the heart attack patients in each of those hospitals.
b. The researchers list all of the heart attack patients and measure the cholesterol level of every 25^{th} person on the list.
c. The researchers go to one hospital on a given day and measure the cholesterol level of the heart attack patients at that time.
d. The researchers list all of the heart attack patients. They then measure the cholesterol levels of randomly selected patients.
e. The researchers divide the heart attack patients based on race, and then measure the cholesterol levels of randomly selected patients in each race grouping.

2. The quality control officer at a manufacturing plant needs to determine what percentage of items in a batch are defective. The following are different sampling techniques that could be used by the officer. Classify each as simple random sample, stratified sample, systematic sample, cluster sample, or convenience sample.

a. The officer lists all of the batches in a given month. The number of defective items is counted in randomly selected batches.
b. The officer takes the first 10 batches and counts the number of defective items.
c. The officer groups the batches made in a month into which shift they are made. The number of defective items is counted in randomly selected batches in each shift.
d. The officer chooses every 15^{th} batch off the line and counts the number of defective items in each chosen batch.
e. The officer divides the batches made in a month into which day they were made. Then certain days are picked and every batch made that day is counted to determine the number of defective items.

3. You wish to determine the GPA of students at your school. Describe what process you would go through to collect a sample if you use a simple random sample.

4. You wish to determine the GPA of students at your school. Describe what process you would go through to collect a sample if you use a stratified sample.

5. You wish to determine the GPA of students at your school. Describe what process you would go through to collect a sample if you use a systematic sample.

6. You wish to determine the GPA of students at your school. Describe what process you would go through to collect a sample if you use a cluster sample.

7. You wish to determine the GPA of students at your school. Describe what process you would go through to collect a sample if you use a convenience sample.

1.3 Experimental Design

The section is an introduction to experimental design. This is how to actually design an experiment or a survey so that they are statistical sound. Experimental design is a very involved process, so this is just a small introduction.

Guidelines for planning a statistical study

1. Identify the individuals that you are interested in. Realize that you can only make conclusions for these individuals. As an example, if you use a fertilizer on a certain genus of plant, you can't say how the fertilizer will work on any other types of plants. However, if you diversify too much, then you may not be able to tell if there really is an improvement since you have too many factors to consider.
2. Specify the variable. You want to make sure this is something that you can measure, and make sure that you control for all other factors too. As an example, if you are trying to determine if a fertilizer works by measuring the height of the plants on a particular day, you need to make sure you can control how much fertilizer you put on the plants (which would be your treatment), and make sure that all the plants receive the same amount of sunlight, water, and temperature.
3. Specify the population. This is important in order for you know what conclusions you can make and what individuals you are making the conclusions about.
4. Specify the method for taking measurements or making observations.
5. Determine if you are taking a census or sample. If taking a sample, decide on the sampling method.
6. Collect the data.
7. Use appropriate descriptive statistics methods and make decisions using appropriate inferential statistics methods.
8. Note any concerns you might have about your data collection methods and list any recommendations for future.

There are two types of studies:

An **observational study** is when the investigator collects data merely by watching or asking questions. He doesn't change anything.

An **experiment** is when the investigator changes a variable or imposes a treatment to determine its effect.

1.3.1 Example: Observational Study or Experiment

State if the following is an observational study or an experiment.

a.) Poll students to see if they favor increasing tuition.

Solution:

This is an observational study. You are only asking a question.

b.) Give some students a tutor to see if grades improve.

Solution:

This is an experiment. The tutor is the treatment.

Many observational studies involve surveys. A **survey** uses questions to collect the data and needs to be written so that there is no bias.

In an experiment, there are different options.

Randomized two-treatment experiment: in this experiment, there are two treatments, and individuals are randomly placed into the two groups. Either both groups get a treatment, or one group gets a treatment and the other gets either nothing or a placebo. The group getting either no treatment or the placebo is called the control group. The group getting the treatment is called the treatment group. The idea of the placebo is that a person thinks they are receiving a treatment, but in reality they are receiving a sugar pill or fake treatment. Doing this helps to account for the placebo effect, which is where a person's mind makes their body respond to a treatment because they think they are taking the treatment when they are not really taking the treatment. Note, not every experiment needs a placebo, such when using animals or plants. Also, you can't always use a placebo or no treatment. As an example, if you are testing a new blood pressure medication you can't give a person with high blood pressure a placebo or no treatment because of moral reasons.

Randomized Block Design: a block is a group of subjects that are similar, but the blocks differ from each other. Then randomly assign treatments to subjects inside each block. An example would be separating students into full-time versus part-time, and then randomly picking a certain number full-time students to get the treatment and a certain number part-time students to get the treatment. This way some of each type of student gets the treatment and some do not.

Rigorously Controlled Design: carefully assign subjects to different treatment groups, so that those given each treatment are similar in ways that are important to the experiment. An example would be if you want to have a full-time student who is male, takes only night classes, has a full-time job, and has children in one treatment group, then you need to have the same type of student getting the other treatment. This type of design is hard to implement since you don't know how many differentiations you would use, and should be avoided.

Matched Pairs Design: the treatments are given to two groups that can be matched up with each other in some ways. One example would be to measure the effectiveness of a muscle relaxer cream on the right arm and the left arm of individuals, and then for each individual you can match up their right arm measurement with their left arm. Another example of this would be before and after experiments, such as weight before and weight after a diet.

No matter which experiment type you conduct, you should also consider the following:

Replication: repetition of an experiment on more than one subject so you can make sure that the sample is large enough to distinguish true effects from random effects. It is also the ability for someone else to duplicate the results of the experiment.

Blind study is where the individual does not know which treatment they are getting or if they are getting the treatment or a placebo.

Double-blind study is where neither the individual nor the researcher knows who is getting which treatment or who is getting the treatment and who is getting the placebo. This is important so that there can be no bias created by either the individual or the researcher.

One last consideration is the time period that you are collecting the data over. There are three types of time periods that you can consider.

Cross-sectional study: data observed, measured, or collected at one point in time.

Retrospective (or **case-control**) **study**: data collected from the past using records, interviews, and other similar artifacts.

Prospective (or **longitudinal** or **cohort**) **study**: data collected in the future from groups sharing common factors.

1.3.2 Homework

1. You want to determine if cinnamon reduces a person's insulin sensitivity. You give patients who are insulin sensitive a certain amount of cinnamon and then measure their glucose levels. Is this an observation or an experiment? Why?

2. You want to determine if eating more fruits reduces a person's chance of developing cancer. You watch people over the years and ask them to tell you how many servings of fruit they eat each day. You then record who develops cancer. Is this an observation or an experiment? Why?

3. A researcher wants to evaluate whether countries with lower fertility rates have a higher life expectancy. They collect the fertility rates and the life expectancies of countries around the world. Is this an observation or an experiment? Why?

4. To evaluate whether a new fertilizer improves plant growth more than the old fertilizer, the fertilizer developer gives some plants the new fertilizer and others the old fertilizer. Is this an observation or an experiment? Why?

5. A researcher designs an experiment to determine if a new drug lowers the blood pressure of patients with high blood pressure. The patients are randomly selected to be in the study and they randomly pick which group to be in. Is this a randomized experiment? Why or why not?

6. Doctors trying to see if a new stint works longer for kidney patients, asks patients if they are willing to have one of two different stints put in. During the procedure the doctor decides which stent to put in based on which one is on hand at the time. Is this a randomized experiment? Why or why not?

7. A researcher wants to determine if diet and exercise together helps people lose weight over just exercising. The researcher solicits volunteers to be part of the study, randomly picks which volunteers are in the study, and then lets each volunteer decide if they want to be in the diet and exercise group or the exercise only group. Is this a randomized experiment? Why or why not?

8. To determine if lack of exercise reduces flexibility in the knee joint, physical therapists ask for volunteers to join their trials. They then randomly select the volunteers to be in the group that exercises and to be in the group that doesn't exercise. Is this a randomized experiment? Why or why not?

9. You collect the weights of tagged fish in a tank. You then put an extra protein fish food in water for the fish and then measure their weight a month later. Are the two samples matched pairs or not? Why or why not?

10. A mathematics instructor wants to see if a computer homework system improves the scores of the students in the class. The instructor teaches two different sections of the same course. One section utilizes the computer homework system and the other section completes homework with paper and pencil. Are the two samples matched pairs or not? Why or why not?

11. A business manager wants to see if a new procedure improves the processing time for a task. The manager measures the processing time of the employees then trains the employees using the new procedure. Then each employee performs the task again and the processing time is measured again. Are the two samples matched pairs or not? Why or why not?

12. The prices of generic items are compared to the prices of the equivalent named brand items. Are the two samples matched pairs or not? Why or why not?

13. A doctor gives some of the patients a new drug for treating acne and the rest of the patients receive the old drug. Neither the patient nor the doctor knows who is getting which drug. Is this a blind experiment, double blind experiment, or neither? Why?

14. One group is told to exercise and one group is told to not exercise. Is this a blind experiment, double blind experiment, or neither? Why?

15. The researchers at a hospital want to see if a new surgery procedure has a better recovery time than the old procedure. The patients are not told which procedure that was used on them, but the surgeons obviously did know. Is this a blind experiment, double blind experiment, or neither? Why?

16. To determine if a new medication reduces headache pain, some patients are given the new medication and others are given a placebo. Neither the researchers nor the patients know who is taking the real medication and who is taking the placebo. Is this a blind experiment, double blind experiment, or neither? Why?

17. A new study is underway to track the eating and exercise patterns of people at different time periods in the future, and see who is afflicted with cancer later in life. Is this a cross-sectional study, a retrospective study, or a prospective study? Why?

18. To determine if a new medication reduces headache pain, some patients are given the new medication and others are given a placebo. The pain levels of a patient are then recorded. Is this a cross-sectional study, a retrospective study, or a prospective study? Why?

19. To see if there is a link between smoking and bladder cancer, patients with bladder cancer are asked if they currently smoke or if they smoked in the past. Is this a cross-sectional study, a retrospective study, or a prospective study? Why?

20. The Nurses Health Survey was a survey where nurses were asked to record their eating habits over a period of time, and their general health was recorded. Is this a cross-sectional study, a retrospective study, or a prospective study? Why?

21. Consider a question that you would like to answer. Describe how you would design your own experiment. Make sure you state the question you would like to answer, then determine if an experiment or an observation is to be done, decide if the question needs one or two samples, if two samples are the samples matched, if this is a randomized experiment, if there is any blinding, and if this is a cross-sectional, retrospective, or prospective study.

1.4 How Not to Do Statistics

Many studies are conducted and conclusions are made. However, there are occasions where the study is not conducted in the correct manner or the conclusion is not correctly made based on the data. There are many things that you should question when you read a study. There are many reasons for the study to have bias in it. Bias is where a study may have a certain slant or preference for a certain result. The following are a list of some of the questions or issues you should consider to help decide if there is bias in a study.

One of the first issues you should ask is who funded the study. If the entity that sponsored the study stands to gain either profits or notoriety from the results, then you should question the results. It doesn't mean that the results are wrong, but you should scrutinize them on your own to make sure they are sound. As an example if a study says that genetically modified foods are safe, and the study was funded by a company that sells genetically modified food, then one may question the validity of the study. Since the company funds the study and their profits rely on people buying their food, there may be bias.

An experiment could have **lurking or confounding variables** when you cannot rule out the possibility that the observed effect is due to some other variable rather than the factor being studied. An example of this is when you give fertilizer to some plants and no fertilizer to others, but the no fertilizer plants also are placed in a location that doesn't receive direct sunlight. You won't know if the plants that received the fertilizer grew taller because of the fertilizer or the sunlight. Make sure you design experiments to eliminate the effects of confounding variables by controlling all the factors that you can.

Overgeneralization is where you do a study on one group and then try to say that it will happen on all groups. An example is doing cancer treatments on rats. Just because the treatment works on rats does not mean it will work on humans. Another example is that until recently most FDA medication testing had been done on white males of a particular age. There is no way to know how the medication affects other genders, ethnic groups, age groups, and races. The new FDA guidelines stresses using individuals from different groups.

Cause and effect is where people decide that one variable causes the other just because the variables are related or correlated. Unless the study was done as an experiment where a variable was controlled, you cannot say that one variable caused the other. Most likely there is another variable that caused both. As an example, there is a relationship between number of drownings at the beach and ice cream sales. This does not mean that ice cream sales increasing causes people to drown. Most likely the cause for both increasing is the heat.

Sampling error: This is the difference between the sample results and the true population results. This is unavoidable, and results in the fact that samples are different from each other. As an example, if you take a sample of 5 people's height in your class, you will get 5 numbers. If you take another sample of 5 people's heights in your class, you will likely get 5 different numbers.

Nonsampling error: This is where the sample is collected poorly either through a biased sample or through error in measurements. Care should be taken to avoid this error.

Lastly, there should be care taken in considering the difference between **statistical significance versus practical significance**. This is a major issue in statistics. Something could be statistically significance, which means that a statistical test shows there is evidence to show what you are trying to prove. However, in practice it doesn't mean much or there are other issues to consider. As an example, suppose you find that a new drug for high blood pressure does reduce the blood pressure of patients. When you look at the improvement it actually doesn't amount to a large difference. Even though statistically there is a change, it may not be worth marketing the product because it really isn't that big of a change. Another consideration is that you find the blood pressure medication does improve a person's blood pressure, but it has serious side effects or it costs a great deal for a prescription. In this case, it wouldn't be practical to use it. In both cases, the study is shown to be statistically significant, but practically you don't want to use the medication. The main thing to remember in a statistical study is that the statistics is only part of the process. You also want to make sure that there is practical significance too. One more comment on statstical significance, the American Statistical Association (ASA) recently came out with a statement, "Based on our review of the articles in this special issue and the broader literature, we conclude that it is time to stop using the term 'statistically significant' entirely." (Advanced Solutions International, Inc, 2019) Though the ASA suggests not using this term anymore, there are many studies that have been done in the past that uses this term, so it is presented here. However, it is not a term that should be use and will be down played in the rest of this book.

Surveys have their own areas of bias that can occur. A few of the issues with surveys are in the wording of the questions, the ordering of the questions, the manner the survey is conducted, and the response rate of the survey.

The wording of the questions can cause **hidden bias**, which is where the questions are asked in a way that makes a person respond a certain way. An example is that a poll was done where people were asked if they believe that there should be an amendment to the constitution protecting a woman's right to choose. About 60% of all people questioned said yes. Another poll was done where people were asked if they believe that there should be an amendment to the constitution protecting the life of an unborn child. About 60% of all people questioned said yes. These two questions deal with the same issue, though giving opposite results, but how the question was asked affected the outcome.

The ordering of the question can also cause **hidden bias**. An example of this is if you were asked if there should be a fine for texting while driving, but proceeding that question is the question asking if you text while drive. By asking a person if they actually partake in the activity, that person now personalizes the question and that might affect how they answer the next question of creating the fine.

Non-response is where you send out a survey but not everyone returns the survey. You can calculate the response rate by dividing the number of returns by the number of surveys sent. Most response rates are around 30-50%. A response rate less than 30% is very poor and the results of the survey are not valid. To reduce non-response, it is better to conduct the surveys in person, though these are very expensive. Phones are the next best way to conduct surveys, emails can be effective, and physical mailings are the least desirable way to conduct surveys.

Voluntary response is where people are asked to respond via phone, email or online. The problem with these is that only people who really care about the topic are likely to call or email. These surveys are not scientific and the results from these surveys are not valid. Note: all studies involve volunteers. The difference between a voluntary response survey and a scientific study is that in a scientific study the researchers ask the individuals to be involved, while in a voluntary response survey the individuals become involved on their own choosing.

1.4.1 Example: Bias in a Study

Suppose a mathematics department at a community college would like to assess whether computer-based homework improves students' test scores. They use computer-based homework in one classroom with one teacher and use traditional paper and pencil homework in a different classroom with a different teacher. The students using the computer-based homework had higher test scores. What is wrong with this experiment?

Solution

Since there were different teachers, you do not know if the better test scores are because of the teacher or the computer-based homework. A better design would be have the same teacher teach both classes. The control group would utilize traditional paper and pencil homework and the treatment group would utilize the computer-based homework. Both classes would have the same teacher, and the students would be split between the two classes randomly. The only difference between the two groups should be the homework method. Of course, there is still variability between the students, but utilizing the same teacher will reduce any other confounding variables.

1.4.2 Example: Cause and Effect

Determine if the one variable did cause the change in the other variable.

a. Cinnamon was giving to a group of people who have diabetes, and then their blood glucose levels were measured a time period later. All other factors for each person were kept the same. Their glucose levels went down. Did the cinnamon cause the reduction?

Solution:

Since this was a study where the use of cinnamon was controlled, and all other factors were kept constant from person to person, then any changes in glucose levels can be attributed to the use of cinnamon.

b. There is a link between spray on tanning products and lung cancer. Does that mean that spray on tanning products cause lung cancer?

Solution:

Since there is only a link, and not a study controlling the use of the tanning spray, then you cannot say that increased use causes lung cancer. You can say that there is a link, and that there could be a cause, but you cannot say for sure that the spray causes the cancer.

1.4.3 Example: Generalizations

a. A researcher conducts a study on the use of ibuprofen on humans and finds that it is safe. Does that mean that all species can use ibuprofen?

Solution:

No. Just because a drug is safe to use on one species doesn't mean it is safe to use for all species. In fact, ibuprofen is toxic to cats.

b. Aspirin has been used for years to bring down fevers in humans. Originally it was tested on white males between the ages of 25 and 40 and found to be safe. Is it safe to give to everyone?

Solution:

No. Just because one age group can use it doesn't mean it is safe to use for all age groups. In fact, there has been a link between giving a child under the age of 19 aspirin when they have a fever and Reye's syndrome.

1.4.4 Homework

1. Suppose there is a study where a researcher conducts an experiment to show that deep breathing exercises helps to lower blood pressure. The researcher takes two groups of people and has one group to perform deep breathing exercises and a series of aerobic exercises every day and the other group was asked to refrain from any exercises. The researcher found that the group performing the deep breathing exercises and the aerobic exercises had lower blood pressure. Discuss any issue with this study.

2. Suppose a car dealership offers a low interest rate and a longer payoff period to customers or a high interest rate and a shorter payoff period to customers, and most customers choose the low interest rate and longer payoff period, does that mean that most customers want a lower interest rate? Explain.

3. Over the years it has been said that coffee is bad for you. When looking at the studies that have shown that coffee is linked to poor health, you will see that people who tend to drink coffee don't sleep much, tend to smoke, don't eat healthy, and tend to not exercise. Can you say that the coffee is the reason for the poor health or is there a lurking variable that is the actual cause? Explain.

4. When researchers were trying to figure out what caused polio, they saw a connection between ice cream sales and polio. As ice cream sales increased so did the incident of polio. Does that mean that eating ice cream causes polio? Explain your answer.

5. There is a positive correlation between having a discussion of gun control, which usually occur after a mass shooting, and the sale of guns. Does that mean that the discussion of gun control increases the likelihood that people will buy more guns? Explain.

6. There is a study that shows that people who are obese have a vitamin D deficiency. Does that mean that obesity causes a deficiency in vitamin D? Explain.

7. A study was conducted that shows that polytetrafluoroethylene (PFOA) (Teflon is made from this chemical) has an increase risk of tumors in lab mice. Does that mean that PFOA's have an increased risk of tumors in humans? Explain.

8. Suppose a telephone poll is conducted by contacting U.S. citizens via landlines about their view of gay marriage. Suppose over 50% of those called do not support gay marriage. Does that mean that you can say over 50% of all people in the U.S. do not support gay marriage? Explain.

9. Suppose that it can be shown to be statistically significant that a smaller percentage of the people are satisfied with your business. The percentage before was 87% and is now 85%. Do you change how you conduct business? Explain?

10. You are testing a new drug for weight loss. You find that the drug does in fact statistically show a weight loss. Do you market the new drug? Why or why not?

11. There was an online poll conducted about whether the mayor of Auckland, New Zealand, should resign due to an affair. The majority of people participating said he should. Should the mayor resign due to the results of this poll? Explain.

12. An online poll showed that the majority of Americans believe that the government covered up events of 9/11. Does that really mean that most Americans believe this? Explain.

13. A survey was conducted at a college asking all employees if they were satisfied with the level of security provided by the security department. Discuss how the results of this question could be biased.

14. An employee survey says, "Employees at this institution are very satisfied with working here. Please rate your satisfaction with the institution." Discuss how this question could create bias.

15. A survey has a question that says, "Most people are afraid that they will lose their house due to economic collapse. Choose what you think is the biggest issue facing the nation today. a) Economic collapse, b) Foreign policy issues, c) Environmental concerns." Discuss how this question could create bias.

16. A survey says, "Please rate the career of Roberto Clemente, one of the best right field baseball players in the world." Discuss how this question could create bias.

References: Advanced Solutions International, Inc. (n.d.). Retrieved July 16, 2019, from https://www.amstat.org/asa/News/ASA-Calls-Time-on-Statistically-Significant-in-Science-Research.aspx