Chapter 1: Collecting and Analyzing Data

Vocabulary

Data: Information in all forms.

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Bias: When data does not represent the population.

Chapter 1 Introduction: The goal of collecting and analyzing data is to understand the world around us. How data is collected is very important. The goal of collecting data is to get "unbiased" data that represents the population. Analyzing biased data may result in incorrect conclusions and lead to a misguided view of the world around us. It is also important to have a goal in mind when you collect data. Are we trying to find a population percentage from categorical data or a population average from quantitative data? Are we trying to show that two variables are related or are we trying to show cause and effect? Data needs to be collected differently depending on what goal you have in mind.

Section 1A – Two Types of Data – Categorical and Quantitative

Vocabulary

Data: Information in all forms.

Analyzing data is an important skill in the modern world. Companies, hospitals, sports teams all need to analyze data in order to make good decisions. But what is data? A good way to think of data is information in all forms. It is often a list of answers to a question or it may be organized in a spread sheet.

One of the most important factors when analyzing data is to determine what type of data you have and how many variables you are analyzing. Let us start with the types of data.

There are two general types of data, categorical and quantitative.

Categorical Data

Categorical data (or qualitative data) are generally labels that tell us something about the people or objects in the data set. For example, what country do they live in, what is the person's occupation, or what kind of pet they have?

Usually categorical data is made up of words (do you smoke - yes or no), but occasionally a number can be used in place of a word. For example, a zip code can be used instead of the place a person lives. The numbers "1" and "2" may be used instead of yes and no. Or the number of the month instead of the name of the month. Notice though it is a number it really represents a word.

Do you Smoke Cigarettes (Yes or No)	What type of Car do you drive?	What Month were you born in?
Yes	Ford	11
No	Honda	2
No	Dodge	5
Yes	Toyota	7
No	Chevy	9
No	Tesla	10
No	Mercedes	1
No	Chevy	6
No	Toyota	3
No	Ford	2

Quantitative Data

Quantitative data are numbers that measure or count something. They usually have units and taking an average makes sense. For example: a list of people's heights in inches, or temperature in degrees Celsius, or a list of how many dogs are there in various animal shelters across Los Angeles. Notice in each of these cases the data is numerical and an average seems appropriate in the context. We can find the average height, the average temperature, or the average number of dogs in animal shelters in Los Angeles.

Height (Inches)	Temperature (Celsius)	Number of Dogs at Animal Shelters in LA
60	21.4	122
65	25	74
68	38.2	68
59	30	39
62.5	29.6	147
73	31.9	26
61.25	36.4	73
70	28	91
64	20.1	31
66	27.5	44

Numbers used as categories

Remember, not all numeric data is quantitative. Ask yourself if the numbers are measuring or counting something and if an average would make sense. For example, averaging a list of the months people are born in would not really tell us anything. In addition, identity numbers like hospital ID numbers, student ID numbers or social security numbers are not measuring anything and an average would not make sense in the context so they are not quantitative.

New Vocabulary

Data: Information in all forms.

Categorical Data: Data consisting of words describing people or objects. Numbers may sometimes be used in place of words.

Quantitative Data: Numerical measurement data, often including units, counts or averages.

Practice Problems Section 1A

1. The American black bear is a medium-sized bear from North America. The following data was taken from 54 American black bears. Classify each column of data as categorical or quantitative. If the column is quantitative, what are the units? If the column is categorical, indicate how many different options there are in that category.

AGE (months)	Month Data Taken	Gender	Head Length (In)	Head Width (In)	Neck Circum (in)	Length (in)	Chest (in)	Weight (Lbs)
19	July	male	11	5.5	16	53	26	80
55	July	male	16.5	9	28	67.5	45	344
81	September	male	15.5	8	31	72	54	416
115	July	male	17	10	31.5	72	49	348
104	August	female	15.5	6.5	22	62	35	166
100	April	female	13	7	21	70	41	220
56	July	male	15	7.5	26.5	73.5	41	262
51	April	male	13.5	8	27	68.5	49	360
57	September	female	13.5	7	20	64	38	204
53	May	female	12.5	6	18	58	31	144
68	August	male	16	9	29	73	44	332
8	August	male	9	4.5	13	37	19	34
44	August	female	12.5	4.5	10.5	63	32	140
32	August	male	14	5	21.5	67	37	180
20	August	female	11.5	5	17.5	52	29	105
32	August	male	13	8	21.5	59	33	166
45	September	male	13.5	7	24	64	39	204
9	September	female	9	4.5	12	36	19	26
21	September	male	13	6	19	59	30	120
177	September	male	16	9.5	30	72	48	436
57	September	female	12.5	5	19	57.5	32	125
81	September	female	13	5	20	61	33	132
21	September	male	13	5	17	54	28	90
9	September	male	10	4	13	40	23	40
45	September	male	16	6	24	63	42	220
9	September	male	10	4	13.5	43	23	46
33	September	male	13.5	6	22	66.5	34	154
57	September	female	13	5.5	17.5	60.5	31	116
45	September	female	13	6.5	21	60	34.5	182
21	September	male	14.5	5.5	20	61	34	150
10	October	male	9.5	4.5	16	40	26	65
82	October	female	13.5	6.5	28	64	48	356
70	October	female	14.5	6.5	26	65	48	316
10	October	male	11	5	17	49	29	94
10	October	male	11.5	5	17	47	29.5	86
34	October	male	13	7	21	59	35	150
34	October	male	16.5	6.5	27	72	44.5	270
34	October	male	14	5.5	24	65	39	202
58	October	female	13.5	6.5	21.5	63	40	202
58	October	male	15.5	7	28	70.5	50	365
11	November	male	11.5	6	16.5	48	31	79
23	November	male	12	6.5	19	50	38	148
70	October	male	15.5	7	28	76.5	55	446
11	November	female	9	5	15	46	27	62
83	November	female	14.5	7	23	61.5	44	236
35	November	male	13.5	8.5	23	63.5	44	212
16	April	male	10	4	15.5	48	26	60
16	April	male	10	5	15	41	26	64
17	May	male	11.5	5	17	53	30.5	114
17	May	female	11.5	5	15	52.5	28	76
17	May	female	11	4.5	13	46	23	48
8	August	female	10	4.5	10	43.5	24	29
83	November	male	15.5	8	30.5	75	54	514
18	June	male	12.5	8.5	18	57.3	32.8	140

2. The following data was taken from various cereals. Classify each column of data as categorical or quantitative. If the column is quantitative, what are the units? If the column is categorical, indicate how many different options there are in that category.

Name	Manufacturer	Target (Adult or Child)	Shelf displayed at store	Calories per serving	Carbs (grams per serving)	Fat (grams per serving)	Fiber (grams per serving)	Potassium (milligrams per serving)	Protein (grams per serving)	Sodium (milligrams per serving)	Sugar (grams per serving)	Vitamins (Percent of daily need per serving)	Consumer Report Magazine Rating	Serving Size (Cups per serving)	Weight (Ounces per serving)
Captain Crunch	Quaker	Child	Middle	120	12	2	0	35	1	220	12	25	18	0.75	1
Cocoa Puffs	General	Child	Middle	110	12	1	0	35	1	180	13	25	23	1	1
Trix	General	Child	Middle	110	13	1	0	25	1	140	12	25	28	1	1
Apple Jacks	Kelloggs	Child	Middle	110	11	0	1	30	2	125	14	25	33	1	1
Corn Chex	Ralston	Adult	Bottom	110	22	0	0	25	2	280	3	25	41	1	1
Corn Flakes	Kelloggs	Adult	Bottom	100	21	0	1	35	2	290	2	25	46	1	1
Nut & Honey	Kelloggs	Adult	Middle	120	15	1	0	40	2	190	9	25	30	0.67	1
Smacks	Kelloggs	Child	Middle	110	9	1	1	40	2	70	15	25	31	0.75	1
Multi-Grain	General	Adult	Bottom	100	15	1	2	90	2	220	6	25	40	1	1
Crackln	Kelloggs	Adult	Top	110	10	3	4	160	3	140	7	25	40	0.5	1
Grape-Nuts	Post	Adult	Top	110	17	0	3	90	3	170	3	25	53	0.25	1
Honey Nut	General	Child	Bottom	110	11.5	1	1.5	90	3	250	10	25	31	0.75	1
Nutri-Grain	Kelloggs	Adult	Top	140	21	2	3	130	3	220	7	25	41	0.67	1.33
Product-19	Kelloggs	Adult	Top	100	20	0	1	45	3	320	3	100	42	1	1
Total Raisin	General	Adult	Top	140	15	1	4	230	3	190	14	100	29	1	1.5
Wheat Chex	Ralston	Adult	Bottom	100	17	1	3	115	3	230	3	25	50	0.67	1
Oatmeal	General	Adult	Top	130	13.5	2	1.5	120	3	170	10	25	30	0.5	1.25
Life	Quaker	Child	Middle	100	12	2	2	95	4	150	6	25	45	0.67	1
Maypo	America	Adult	Middle	100	16	1	0	95	4	0	3	25	55	1	1
Quaker Oats	Quaker	Adult	Top	100	14	1	2	110	4	135	6	25	50	0.5	1
Muesli R	Ralston	Adult	top	150	16	3	3	170	4	150	11	25	34	1	1
Quaker Oatmeal	Quaker	Adult	Bottom	100	14	2	2.7	110	5	120	0	0	51	0.67	1
Cheerios	General	Child	Bottom	110	17	2	2	105	6	290	1	25	51	1.25	1
Special K	Kelloggs	Adult	Bottom	110	16	0	1	55	6	230	3	25	53	1	1

3. Determine if the each of the following variables are quantitative or categorical.

a) The number of milligrams of Aspirin given to heart attack patients.
b) The various types of cars being sold at a used car lot.
c) Determining if a person smokes marijuana or not.
d) The number of bicycles sold at various bicycle stores in Seattle, WA.
e) The types of birds observed in Florida.
f) The number of grams of gold found in various streams across northern California.
g) The various types of cardio classes offered at gyms across Los Angeles, CA.
h) The number of cardio classes offered at gyms across Los Angeles, CA.
i) The city a person lives in.
j) The amount of money in peoples' bank accounts.
k) The various zip codes from addresses at a post office.
l) The drivers' license numbers from various taxi drivers.
m) The number of taxis driven in New York City on various days of the week.

Section 1B – Collecting Data

Vocabulary

Data: Information in all forms

Categorical Data: Data consisting of words describing people or objects. Numbers may sometimes be used in place of words.

Quantitative Data: Numerical measurement data, often including units, counts or averages.

Population: The collection of all people or objects you want to study.

Census: Collecting data from everyone in the population.

Sample: Collecting data from a small subgroup of the population.

Bias: When data does not reflect the population.

Random: When everyone in the population has a chance to be included in the sample.

Sample Size: The total number of people, animals or objects you collect data from.

Sampling Bias: Using a bad method to collect data like convenience or voluntary response. Not incorporating randomization in your sample.

One of the most important goals in data science is to learn about the world around us (populations). A population is the collection of all people or objects you want to study. It is very difficult to understand populations sometimes because data may be biased and not reflect the population very well. Bias can occur in many different ways, but certain ways people collect data have more bias than others do. Using a method for collecting data that increases bias is sometimes called "sampling bias". It is important to be aware of various methods used to collect data, the good and the bad.

Method 1: Census

A census is the best way to collect data if it is possible. If our goal is to learn about the population, it makes sense to collect data from everyone in the population. There are ways for a census to be biased, but in terms of the collecting method, a census is the best. Unfortunately, it is almost impossible to collect a census if your population is large. Most statisticians and data scientists are only able to collect a sample, data collected from a small subgroup of the population.

Method 2: Simple Random Sample

If a statistician or data scientist cannot collect a census, the preferred method is to collect a random sample. A random sample is one where everyone in the population has a chance to be in the sample, so it tends to represent the population better than other non-random samples. It is nowhere near as good as a census, but as I said, a census is usually not possible.

We should probably start with discussing the word "random". "Random" in data science and statistics is used very differently than the way people generally use the word. Selecting data randomly means that everyone in the population has a chance to be included in the sample data. You are not collecting data from the millions of people or objects in the population. That would be a census. You are still collecting a sample (small subgroup of the population), but everyone of the millions in the population has a chance to be included. Random samples are often difficult to set up.

Example: A person may walk into a store and say "I chose a person randomly to talk to". They mean that they talked to whoever they ran into. This is not random in statistics. Not everyone of the millions in your population has a chance to be in that store and bump into the person.

A simple random sample is the most common type of random sample. In a simple random sample, individuals in the population are selected randomly. This can be a difficult process. The usual method is to assign everyone in a population a number and then use a random number generator in a computer program to pick random numbers. Computer programs have many built in randomization functions for this purpose. If you have a spreadsheet of the entire population, a computer can also randomly select individuals from the list. The key with a "simple random sample" is that you are selecting people or objects one at a time. Collecting data randomly and one at a time gives greater flexibility to your sample. Almost any grouping is possible with a simple random sample, so it tends to represent populations better than other samples.

There are many examples of a simple random sample. Many statistics companies use a random phone number generator that randomly gives phone numbers. They then call the phone numbers randomly chosen and try to get information from people that answer the phone. The U.S. government may have a computer randomly select social security numbers to select individuals for a sample. A company may have a computer randomly select employee ID numbers to select individuals for a sample.

Method 3: Convenience Sample

People often find collecting a census or a simple random sample difficult, so they chose to collect data in whatever way seems easiest. A sample collected this way is often called a "convenience sample" and is popular with people not trained in statistics. A convenience sample usually has much more bias than a random sample and may not represent the population very well.

An example of a convenience sample is collecting data from your friends and family. This is fine if your population of interest is your friends and family, but will by no means represent a large population. Another example might be standing outside of a store or post office and collecting data from people that leave the store. Beginning statistics students may walk into a mall and collect data from whomever they bump into. They mistakenly think that these are random samples, but they are not. A random sample means everyone in the population has a chance to be included in the sample. Not everyone in the population has a chance to bump into you at a mall or come out of a store at 2:30 pm on a Tuesday afternoon. These are convenience samples and generally do not reflect the population very well.

Method 4: Voluntary Response Sample

Some say that all surveys are bad, but that is not the case. A survey is just a form to collect data from people. When a company takes a census of all its employees, it may require all of the employees to fill out a survey. That is a census. As long as no other forms of bias creep into the data, a census will probably be a very good representation of the population. The point is that giving a survey is not the issue. The issue is whom you give the survey to and who is allowed to fill out the survey.

A voluntary response sample puts a survey out into the world and allow anyone to respond. The usual method used today is to put a survey on a website and allow anyone that comes across the survey to answer. The survey can also be a mailed to every address in a given population. Again, those that fill it out self-select themselves to be in our data.

On the surface, a voluntary response sample may seem like a good way of collecting data. It usually gives a large amount of data. Does this really allow everyone in the population a chance to answer? It turns out the answer is no. Ask yourself the following question. When you are surfing the web and a survey pops up, do you fill it out? I have been asking my statistics classes that question for years and rarely have anyone that says that they do fill out surveys. The key problem is that only certain types of people will fill out a survey voluntarily. It may be a person who is bored and has nothing better to do. It is certainly not a person with three children, working a full time job and going to college full time. It may also be a person who is upset by or feels very passionate about the topic in the voluntary response survey. They are so upset by the lack of pay for teachers that they are willing to fill out a survey to tell you

what they think. The point is that voluntary response surveys tend to over-sample people that are bored or upset and under-sample everyone else. For this reason, voluntary response samples can be very biased and may not represent the population very well.

Note about sample size:

Students often ask me about the importance of how many people or objects you collect data from. This is called "sample size". More data is usually better. A simple random sample of 250 people is better than a simple random sample of 50 people. Is a voluntary response sample of five thousand people better than a random sample of fifty people?" I would tell them that though sample size is important, method is important also. The voluntary response sample of five thousand would tend to over-represent people that are bored or upset about the topic. It does not represent typical people in the population. The random sample of fifty people, while a small sample size, at least does not have that bias.

Summary

So let us summarize the various methods.

- An unbiased census is the best way to collect data to represent a population, because we are collecting data from everyone in the population. An unbiased census is generally better than a random sample.
- If you cannot do a census, then use an unbiased random sample. A simple random sample is most common. The main thing is that if you are collecting a sample, randomization needs to be involved. Random means that everyone in the population has to have a chance to be included in the sample.
- Voluntary response samples and convenience samples tend to be very biased and should be avoided if possible.

Practice Problems Section 1B

Directions: For each of the following, identify the population of interest. Then identify the method used to collect the data (census, convenience, voluntary response, or simple random). Explain why you chose your answer. Was this a good or bad way to collect the data?

1. The admissions department at a college wants to see how many of their students would be in favor of using a new program to register for classes. They put a link on their website so that any students that want to try out the program can. The students can then take a survey and say how well they like the new system.

a) What was the population of interest?
b) What method was used to collect the data? Explain.
c) Was this a good or bad way to collect the data?

2. Michelle, a teacher at a local high school, wants to see how many students at her high school will be attending community college. She gives the students in her one section of advanced placement U.S. History a questionnaire to fill out that asks where they will be attending college.

a) What was the population of interest?
b) What method was used to collect the data? Explain.
c) Was this a good or bad way to collect the data?

3. Jamie is working at the Republican recruiting committee in her city. She is curious how many people that live in her city will vote for the Republican candidate in the next election. She uses a computer to randomly select phone numbers in her city. She then calls those phone numbers to ask people about their voting preferences.

a) What was the population of interest?
b) What method was used to collect the data? Explain.
c) Was this a good or bad way to collect the data?

4. Micah is the CEO of large software development company. He wants to see if his employees have any ideas about areas of software development that the company should pursue. He has every single employee in his company fill out a questionnaire outlining his or her ideas. He gives the employees a stipend on their paycheck to pay them for their time it took to fill out the questionnaire.

a) What was the population of interest?
b) What method was used to collect the data? Explain.
c) Was this a good or bad way to collect the data?

5. Tara wants to collect data on people living in Portland Oregon. She wants to know how many cups a coffee they drink per day. She went to a few supermarkets close to her house and asked people as they were leaving the store.

a) What was the population of interest?
b) What method was used to collect the data? Explain.
c) Was this a good or bad way to collect the data?

6. Julius works for a company in Toronto, Canada that manufactures eyeglasses. He wants to know what styles of glasses people in Toronto prefer. He randomly selects phone numbers in Toronto and calls them to ask about glasses preference.

a) What was the population of interest?
b) What method was used to collect the data? Explain.
c) Was this a good or bad way to collect the data?

7. Hugo works at a public library and wants to collect data on all of the people that come to the library. He looks up every single person in the library database and notes the number of books that he or she has checked out in the last six months.

a) What was the population of interest?
b) What method was used to collect the data? Explain.
c) Was this a good or bad way to collect the data?

8. A company is designing a new type of smart phone. They want to know how much memory people prefer in their smart phones. The put a question up on several search engines and allow anyone to answer.

a) What was the population of interest?
b) What method was used to collect the data? Explain.
c) Was this a good or bad way to collect the data?

9. A college wants to collect data on their students to see how often they use the various student services offered by the college. They randomly select 50 student ID numbers and collected data from all of the students chosen.

a) What was the population of interest?
b) What method was used to collect the data? Explain.
c) Was this a good or bad way to collect the data?

Section 1C – Bias

Vocabulary

Data: Information in all forms

Categorical Data: Data consisting of words or numbers used in place of words.

Quantitative Data: Numerical measurement data.

Population: The collection of all people or objects you want to study.

Census: Collecting data from everyone in the population.

Sample: Collecting data from a small subgroup of the population.

Bias: When data does not reflect the population.

Random: When everyone in the population has a chance to be included in the sample.

Sample Size: The total number of people, animals or objects you collect data from.

Sampling Bias: Using a bad method to collect data like convenience or voluntary response. Not incorporating randomization in your sample.

The purpose of collecting data is to learn about the world around us, to learn about populations. The problem is that many people that collect data may not have had any training in Statistics or Data Science. The result is that many data sets collected do not reflect the population very well. When this happens, we say that the data is biased.

Many people think that if you collect a random sample or a census, it will guarantee that you will have an unbiased data set. This is not true. There are many types of bias and it is possible to have a census or a random sample that does not reflect the population very well. It is critical that we be aware of these other forms of bias and to try our best to make sure they are not incorporated into our data sets.

Sampling Bias

In the last section, we said that the best way to collect data is a census. This means that we collected data from everyone in the population. If we cannot collect a census then we should try to collect a random sample or at least a sample that represents the population. We said that convenience samples or voluntary response samples are inherently biased and usually do not reflect populations very well. Using a bad data collecting method like convenience or voluntary response gives rise to sampling bias. When sampling bias occurs, it usually means the technique for collecting the data was poor.

Question Bias

It has been said that there are lies, bad lies, and then there is statistics. There is some truth in this. People with specific agendas may twist data and statistical analysis to suit their purpose. One way to do this is question bias.

A question bias occurs when someone phrases a question in a specific way to force people to answer the way they want.

For example, suppose a politician wants to show that most people in her city agree with her policy on raising taxes to improve health care. She may collect a great simple random sample, but ask the question this way.

"Health care in our city is extremely bad. Hospitals and urgent cares are in bad need of renovation and need better supplies. The elderly need to know that we have not forgotten them. We need to improve the quality of care for our children. Will you support my policy for improving health care across our city?"

Phrasing the question this way, no one would guess that the real issue was whether to raise taxes. People, hearing this question, think about helping the children and elderly, not about taxes. When a large percentage of people answer that they support her plan, she now has data to support her agenda.

When you collect data, you want to ask questions in a neutral way that does not attempt to sway people in one direction or another. It also should not leave out key information like what the real question is. If the politician had simply asked people in the simple random sample if they would be in favor of raising taxes to improve health care, she likely would have gotten a much smaller percentage of people to agree.

Notice that in this example, the data was a simple random sample. This is a good data collection method, as methods go. However, the incorporation of a question bias into the data makes the data very bad. This simple random sample does not reflect the population at all. The data has been manipulated to support an agenda.

Response Bias

Many topics are very difficult to get data on because people do not feel comfortable answering truthfully. If you ask people if they are addicted to alcohol or drugs, they are likely to deny it even if they do struggle with substance addiction. People may lie about their age, weight, or salary. When a large percentage of people in your data lie, you have a response bias in your data.

Suppose a church wants to collect data on how many hours per week their congregation spends helping the homeless. They decide to have every person in their congregation fill out a survey listing how many hours per week they help the homeless. Remember a census is usually the best way to collect data about a population, but this census has a problem. It is a topic that people are likely to lie about. People may put a higher number of hours on the survey than they really do so that they will not look bad to the church leaders. The average number of hours calculated from this data will likely be larger than the population average number of hours. Even though this is a census, it probably does not reflect the population very well.

When dealing with topics that people are likely to lie about, the data scientist needs to have a plan to deal with the response bias. Instead of asking people their weights, maybe they weigh them on a scale. Instead of asking people about their salary, maybe they look at paycheck stubs. Instead of asking people about substance abuse, they may collect data from agencies that support people with addiction.

Deliberate Bias

We have stated already that people may misuse statistics and data in order to support their agenda. Deliberate bias is another example of this. Deliberate bias can take on a variety of forms. It could be someone deliberately leaving out groups from the data. The most common is collecting data and then leaving out the data of people that disagreed with you. It can also be deliberately lying about the results of the data report. Maybe the data makes your restaurant or hospital or school look bad, so people just falsify their records and deliberately lie about the results of the study. The data may be census or a random sample put the conclusions have been falsified and the data distorted.

Deliberate bias is a major problem in statistics. It is also a good reason to have an independent statistics company collect the data and do the analysis. Use a statistics company that is not tied to the government, business, hospital, restaurant or politician in question. An independent statistics company is less likely to lie about the results or to falsify the data, though it is naive to think that it never happens.

I tend to be suspicious about internal statistics reports that come out where the company, government or politician refuses to share the data. We are supposed to take their word for it and agree with the findings. There are good reasons why companies do not share data, but I always wonder if they are they afraid that someone analyzing that data would come to a very different conclusion?

There is large worldwide discussion of ethics for people that work in the fields of statistics or data science. Statistical analysis is a powerful tool and is a vital discipline to understand and improve the world around us, but falsifying records or manipulating data should never be an option. It is not only unethical, but also makes people question the integrity of our science.

Sometimes specific groups in the population may not be represented very well in the data. This also falls under the umbrella of deliberate bias. For example, suppose a person may wish to collect data on adults living in a city. However, they only collected data from people living in the wealthier areas of that city. It may not have been done deliberately. It could just be that the person collecting the data did not think about certain groups in the population that are not being represented. In large cities, the homeless are often difficult to get data on. A person collecting data has to have a plan for getting data that will represent all the groups in their population, including the homeless.

Non-response Bias

Non-response bias is becoming a huge problem for all people that collect data. A computer may randomly select people to collect data from, but more often than not, the person does not want to participate. They may fear identity theft or are just too busy to participate. It is a huge problem. We need data. We need to understand the world around us, but it now becoming increasing difficult to get unbiased data. Many people that collect data report that sometimes only one in every five randomly selected people will participate and give data. The problem of non-response bias continues to get worse. This makes us consider what type of person gives data and if that person is truly reflective of all people in the population.

To combat the problem of non-response bias, many people that collect data offer a reward system for people that will participate and give data. This may help a little, but then offering a reward may incorporate its own bias into the data.

Summary

There are many reasons why data may not reflect a population. It is a mistake to think that a random sample or a census will always be devoid of bias. It is increasingly important to be aware of possible sources of bias and to strive to keep them out of our data as much as possible. The goal of data collecting is to collect unbiased data that reflects the population. Always phrase questions in a neutral way that avoids question bias. Have a plan for collecting data about topics where people are likely to lie. We have to have a good plan on how we will collect data. It should be a census or a random sample, but we should also think about groups that may not be represented. We need to avoid deliberate bias and never falsify reports or distort data to support someone's agenda.

Practice Problems Section 1C

1. Define each of the following and give an example of each.

    a) Population
    b) Census
    c) Sample
    d) Bias
    e) Question Bias
    f) Response Bias
    g) Sampling Bias
    h) Deliberate Bias
    i) Non-response Bias

Directions for #2-8: For each of the following scenarios, describe the population of interest and all of the types of bias that the data may have (Question, Response, Sampling, Deliberate or Non-response). There may be more than one type of bias involved. Explain your answers.

2. We are interested in finding what percent of people in the U.S. agree or disagree with vaccinating children. To figure this out, we randomly selected 350 people in the U.S. and asked them the following question: "In order to save children from devastating diseases, do you agree that all children should be vaccinated?"

3. We are interested in finding out what percent of Americans use Cocaine. We randomly chose 400 Americans and asked them if they use Cocaine or not.

4. What is the average age of college students in Canada? Since my cousin lives in Canada, I asked him to drive to two colleges near his house and ask people he bumps into what their age is.

5. Julie is interested in calculating the yearly income of adults in Palmdale. She drives around Palmdale, stops at certain streets, and then asks people that live on that street what their yearly income is? She skips streets that look "sketchy" as she is worried about her safety.

6. A college wants to collect data on their students to see how often they use the health office for mental health counseling. They took a simple random sample of college students and asked the following question. "It is very important for all college students to have mental health support. College students report having depression, anxiety and high stress levels. The college offers free mental health counseling at the health office. Have you taken advantage of these mental health services?"

7. A pharmaceutical company took random samples of their pills to check that the pill has the correct type and amount of medicine. They noticed that several of their pills did not have the correct amount of medicine, but decided to delete this data.

8. An auto manufacturer wants to collect data on the type and number of mechanical problems in their cars. They decide to keep data only on all cars brought to their dealerships nationwide.

9. A computer algorithm called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) was created by Northpointe, Inc. The algorithm assesses whether defendants have a higher or lower risk of repeating crimes. Northpointe, Inc. did the validation study to show that the algorithm works.

Section 1D – Experimental Design

Vocabulary

Explanatory Variable: The independent or treatment variable. In an experiment, this is the variable causes the effect.

Response Variable: The dependent variable. In an experiment this the variable that measures the effect.

Confounding Variables (or lurking variables): Other variables that might influence the response variable other than the explanatory variable being studied.

Experimental Design: A scientific method for controlling confounding variables and proving cause and effect.

Random assignment: Take a group of people or objects and randomly split them into two or more groups. This creates similar groups and helps control confounding variables.

Placebo: A fake medicine or fake treatment used to control the placebo effect.

Placebo Effect: The capacity of the human brain to manifest physical responses based on the person believing something is true. A placebo (fake medicine) is often given to control the placebo effect.

In statistics, we often want to determine if there is a relationship or association between two variables. We also may want to measure the strength of the relationship. For example, we may want to know if there is a relationship between blood pressure and heart rate. We may want to see if living in tropical climates is associated with having nut allergies.

In order to show that two variables are related or associated we use an observational study. We would collect data and use statistical methods to analyze and measure the strength of the relationship. However, showing that two variables are related does not prove that one causes the other.

Association ≠ Causation!!!!

Why?

Let us suppose that we have shown that there is a strong relationship between drinking alcohol and getting into a car accident. This tells us that alcohol consumption is an important factor to be considered when studying car accidents. However, this does not prove that drinking alcohol causes car accidents. Many factors go into having a car accident besides how much alcohol they consume. Can you name a few?

Other factors that may influence having a car accident besides alcohol: age of driver, experience of the driver, condition of the car, traffic, road conditions, weather, other drivers, distractions (like texting, eating or changing a radio station), using drugs, …

These are called “confounding variables”. Confounding variables are factors that might influence your response variable other than the explanatory variable you are studying. In this case, factors that might influence having a car accident other than how much alcohol the driver consumed. Some statistics books call these “confounding variables” or “lurking variables”.

Note: The explanatory variable (alcohol consumption) is not a confounding variable. Alcohol is the explanatory variable we were studying. Confounding variables are factors other than alcohol that might influence the response (car accident).

Here is the point. If many variables were involved in having a car accident, it would be wrong to say that the alcohol was solely responsible for the car accident. Alcohol is just one of many factors involved. We have shown that drinking alcohol is related but we have not proven cause and effect. To prove cause and effect we need to deal with the confounding variables.

Experimental Design

So how do we prove cause and effect? It is difficult. You would need to prove that each confounding variables is not involved and so it is only the explanatory variable that is causing the response. The key is controlling the confounding variables. Thankfully, scientists have put a great deal of thought into this process of controlling confounding variables and proving cause and effect. We call this process "experimental design".

Experimental design is a scientific method for controlling confounding variables and proving cause and effect. A key component to experimental design is the creation of similar groups through random assignment.

To control confounding variables, we will need to create two or more groups of people or objects that are very alike. One way to do this is by "random assignment". Random assignment is a process where you take a group of people or objects and randomly split them into two or more groups. The randomly assigned groups tend to be very similar. If we do not think the groups are similar enough, we can use techniques like blocking or direct control to make the groups even more alike.

Another way to make alike groups is to use the same group of people twice. Think about it. The two groups would be perfectly alike. They would have the same ages, same amount of stress, same genetics, same blood pressures and the same jobs.

Example

Let us look at the previous example. How do we prove that drinking alcohol does cause car accidents?

Explanatory (treatment) Variable: Drinking alcohol or not

Response Variable (what we will measure): Did the person get into a car accident or not?

So how do we set up an experiment to prove that drinking alcohol causes car accidents? The first thing is to list out your possible confounding variables.

Possible Confounding Variables: age of driver, experience of the driver, condition of the car, traffic, road conditions, weather, other drivers, distractions (like passengers, texting, eating or changing a radio station), other drugs, gender, race, genetics, reflexes.

To control the confounding variables, we need to create two groups of people. The two groups should be the same (or at least as similar as possible) in all areas that the confounding variables address. Therefore, the groups should have similar ages, similar driving experience, similar cars and car condition, similar road conditions and similar distractions, similar genders, similar race and ethnicity, similar genetics and reflexes.

There are two ways to go about this. Let us suppose we have a group of 80 adult paid volunteers to conduct this experiment. One option would be to randomly put the volunteers into two groups and try to make the groups as similar as possible. A better option in this case would be to use the same people twice.

We had the people in the experiment drive an obstacle course sober. They must have no alcohol or other drugs in their system. They all used the same car on the same track with the same weather. The course was designed with cones and we will monitor how many cones the people hit. They all were not allowed to have any other person in the car. There was no other distractions as radios and phones were not allowed. We will monitor how many car accidents they had by checking how many cones they hit.

Now we will have all the people drink a certain amount of alcohol and then drive the course again. It is important to see that the alcohol (treatment) group was made up of exactly the same people as the sober (control) group. The response variable we measured was the number of cones they hit.

Conclusion

The results found that the alcohol group hit significantly more cones (significantly more car accidents) than the sober group. We have now proven that drinking alcohol causes car accidents.

Think about it. It cannot be the ages of the drivers or driving experience. The two groups had the exact same ages and the exact same driving experience. It cannot be gender, race, genetics, or reflexes. The two groups had the exact same genders, race, genetics, and reflexes. It cannot be drugs or other distractions like phones or radios. Neither group had drugs or any other distractions. If you notice, every one of the confounding variables is the same in the two groups. The only difference was that one group had alcohol and the other did not. Therefore, the only reason why the alcohol group had significantly more accidents is the alcohol. The experiment has proven that drinking alcohol causes car accidents.

Note: It is easy to confuse the two variables in an experiment with the two groups. They are not the same thing.

In this case, the explanatory variable is having alcohol or not. The response variable is the number of cones (accidents) the drivers had. The two groups are decided by those that have explanatory variable (alcohol) and those that do not. In this case, the two groups are the exact same people measured twice.

We usually call the group that has the explanatory variable the "treatment group" and the group that does not have the explanatory variable the "control group".

Example 2

When a pharmaceutical company needs to prove that a medicine works, they must use experimental design. In the United States, pharmaceutical companies have to prove to the Food and Drug Administration (FDA) that their medicine has the effect it is supposed to and is relatively safe with few side effects.

Suppose a company has a new blood pressure medicine on the market and needs to prove to the FDA that taking it does decrease a person's blood pressure. The company needs to prove cause and effect.

If we have to prove cause and effect, we need an experiment. The first step is to think about the possible confounding variables. What are some reasons why a person's blood pressure might decrease other than taking this new medicine?

Possible Confounding Variables? Stress, Diet, Exercise, Genetics, Age, Gender, Race, Genetics, taking other medicines …

To set up the experiment we need to create two groups of people that are similar in these areas. We start with a group of volunteers with high blood pressure that want to try out this new medicine. We randomly assign the people into two groups. Amazingly when scientists randomly assign people into two groups, the groups tend to be a lot alike. The two groups would have similar numbers of people in each race, similar number of males and females, similar numbers of stressed out people, similar numbers of people that exercise a lot or do not exercise. The people running the experiment can also exercise direct control and intentionally assign people to certain groups to make the groups even more alike.

Human Brain (placebo effect)

There is a problem with our experiment. If a person believes something is true, their brain can tell the body to manifest physical responses. We call this the "placebo effect". Think of it this way. The group that thinks they are getting blood pressure medicine will not be as stressed out about it and their blood pressure may decrease slightly because of that belief. Similarly, the group that thinks they are not getting blood pressure medicine will be more stressed and worried and their blood pressure may increase because of that belief. In a sense, the human brain is a confounding variable that we need to control.

Placebo (fake medicine)

To control the placebo effect as a confounding variable, we need the groups to believe the same thing. One group cannot think they are getting medicine, and the other group cannot believe they are not getting medicine. So we introduce a placebo or fake medicine. The treatment group gets the real blood pressure medicine and the control group gets a fake medicine (placebo). No one in the experiment knows if he or she will be receiving real medicine or a placebo. Some may ask, "Won't that make them more stressed and increase their blood pressure?" Yes. The key is that the two groups will be equally stressed and believe the same thing. That way we control the placebo effect.

For this to work, the people in the experiment cannot know if they are getting the medicine or a placebo. This is called "single blind". When scientists first started using placebos, they were shocked to find that the people in the experiments somehow knew if it was a placebo. This defeated the whole purpose. It turned out they could tell by the body language of the person giving the medicine. The person giving the medicine tended to act differently if they were giving the real medicine verses a placebo. So the standard for an experiment about medicines is to use a "double blind" approach. A double blind experiment means that neither the people in the experiment, nor the people giving the medicine, know if it is a placebo or not. Someone knows though. The scientists keep very careful track of who receives a placebo and who receives the medicine. The person directly giving the medicine or placebo cannot know if it is a placebo or not.

Double blind works well. The people in the experiment no longer know if they are receiving a placebo or the real medicine. The experimental design has controlled the placebo effect.

Conclusion

Since we have controlled all of the confounding variables, the experiment has the possibility of proving cause and effect. We still need to see the blood pressures of both groups and make a conclusion. If the treatment group had a significantly lower average blood pressure than the control group, this would prove that taking the medicine does cause a person to have lower blood pressure. If the treatment group and control group have relatively the same average blood pressure, then we may conclude that the medicine is not effective in lowering blood pressure. This would be bad news for the pharmaceutical company. Deciding if one group is significantly higher than another can be very difficult. We will study confidence intervals, test statistics and P-value in later chapters to address this.

Summary

Use an experiment to control confounding variables and prove cause and effect. The groups in the experiment should be the same people either measured multiple times or separated by random assignment. The main idea is that the groups should be very similar in all areas that involve confounding variables. Experiments with medicines should be double blind with a placebo to control the placebo effect.

Use an observational study to see if there is a relationship (association) between two things. Remember observational studies do not control confounding variables, so cannot prove cause and effect.

How can I tell if a study is an experiment or not? Generally, look for random assignment. An experiment usually does not have a random sample of people from the population. The people in the experiment are usually volunteer. The volunteers are then randomly assigned into two or more groups. Random assignment means that they are not trying to apply something to the population, but instead are trying to use experimental design in order to prove cause and effect. If a study takes a random sample from the population, but does not randomly assign, it is probably just an observational study and cannot prove cause and effect.

Note: It should be noted that there are more complex forms of experiments than the types listed in this section. It may not be possible to randomly assign people into two groups. In that case, the scientist need to prove that each confounding variable is not involved. That is a more complex case that you may see in more advanced statistics classes.

Practice Problems Section 1D

(#1-10) Define the following terms.

1. Observational Study

2. Experiment

3. Explanatory Variable

4. Response Variable

5. Confounding Variables

6. Random Assignment

7. Placebo

8. Placebo Effect

9. Single Blind

10. Double Blind

(#11-12) Directions: Answer the following questions about the experiments described.

11. College students in the United States have long claimed that listening to music while studying causes them to retain information at a higher rate. We want to prove that this is not true. Listening to music while studying does not cause a person to retain information at a higher rate. We took a group of volunteer college students and randomly put them into three groups. The people in each group had to memorize the same information. They were then ranked as high retention or low retention. One group had to listen to their favorite music, another group had to listen to a music they hated, and the third group had no music at all. The volume of music was the same for all of the people.

    a) Was random assignment used in the experiment?

    b) List as many confounding variables as you can for this experiment?

    c) What is the explanatory variable (cause) and the response variable (effect) in this experiment?

    d) Describe the treatment group and the control group. Were they alike in the confounding variables?

    e) Describe how the confounding variables were controlled.

    f) The results of the experiment were that the hated music group and the liked music group did about the same. Both music groups did much worse than the no music group. The no music group had significantly better retention than either of the music groups. Does this prove that listening to music does not cause a person to memorize information better? Why or why not?

12. Dramamine is a common medication used in preventing and treating nausea, vomiting and dizziness caused by motion sickness. This medication has become a staple for thousands of people who travel by boat, car or plane. We need to prove that Dramamine is effective in preventing and treating the symptoms of motion sickness. Volunteers were randomly assigned into two groups. One group received Dramamine and the other received a placebo. The amount of motion was the same for all of the people. They were then asked to rank their motion sickness on a scale of 1 to 10.

a) Was random assignment used in the experiment?

b) List as many confounding variables as you can for this experiment?

c) What is the explanatory variable (cause) and the response variable (effect) in this experiment?

d) Describe the treatment group and the control group. Were they alike in the confounding variables?

e) Describe how the confounding variables were controlled.

f) If the Dramamine group has significantly less motion sickness than the placebo group, does this prove that taking Dramamine causes a person to have less motion sickness? Why or why not?

13. An experiment was done on labor market racial discrimination. Statisticians created fictitious resumes to help-wanted adds in Boston and Chicago newspapers. Resumes were randomly assigned to either have a very African American sounding name or a very white sounding name. The results that the percentage of callbacks for resumes with white names was significantly higher than for African American names.

a) Was random assignment used in the experiment?

b) List as many confounding variables as you can for this experiment?

c) What is the explanatory variable (cause) and the response variable (effect) in this experiment?

d) Describe the treatment group and the control group. Were they alike in the confounding variables?

e) Describe how the confounding variables were controlled.

f) Does this experiment prove that there is racial discrimination against African Americans when applying for a job in Boston and Chicago? Why or why not?

Chapter 1 Review Sheet

Key Vocabulary Terms

Data: Information in all forms.

Categorical data: Also called qualitative data. Data in the form of labels that tell us something about the people or objects in the data set. For example, the country they live in, occupation, or type of pet.

Quantitative data: Data in the form of numbers that measure or count something. They usually have units and taking an average makes sense. For example, height, weight, salary, or the number of pets a person has.

Population: The collection of all people or objects to be studied.

Census: Collecting data from everyone in a population.

Sample: Collecting data from a small subgroup of the population.

Random: When everyone in the population has a chance to be included in the sample.

Simple Random Sample: Sample data in which individuals are selected randomly. This method tends to minimize sampling bias and is generally considered a good way to collect data.

Convenience Sample: Sample data that is collected in a way that is easy or convenient. This method tends to have a significant amount of sampling bias and is generally considered a bad way to collect data.

Voluntary Response Sample: Sample data that is collected by putting a survey out into the world and allowing anyone to fill it out. This method tends to have a significant amount of sampling bias and is generally considered a bad way to collect data.

Bias: When data does not represent the population.

Sampling Bias: A type of bias that results from collecting data without using a census or random sample. The method of collecting is flawed. For example, using convenience or voluntary response method to collect the data. We can minimize this bias by collecting the data with a census or random sample.

Question Bias: A type of bias that results when someone phrases the question or gives extra information with the goal of tricking the person into answering a certain way. We can minimize this bias by phrasing our questions in a neutral way and not attempt to sway the person giving data.

Response Bias: A type of bias that results when people giving the data do not answer truthfully or accurately. To minimize this bias, we should collect the data anonymously and assure the person giving the data that the data will be used for scientific purposes and will not be released.

Non-response Bias: A type of bias that results when people refuse to participate or give data. To minimize non-response bias, you may give an incentive like a gift card to encourage people to give data.

Deliberate Bias: A type of bias that results when the people collecting the data falsify the reports, delete data, or decide to not collect data from certain groups in the population. To minimize deliberate bias, the people collecting and analyzing the data need to have good ethics. They should not falsify reports, delete data or leave out groups from the population.

Experimental Design: A scientific method for controlling confounding variables and proving cause and effect.

Observational Study: Collecting data without controlling confounding variables. This type of data cannot prove cause and effect.

Explanatory Variable: The independent or treatment variable. In a cause and effect experiment, this is the cause variable.

Response Variable: The dependent variable. In a cause and effect experiment, this the variable that measures the effect.

Treatment Group: The group of people or objects that has the explanatory variable. In an experiment involving medicine, this would be the group that receives the medicine.

Control Group: The group of people or objects that is used to compare and does not have the explanatory variable. In an experiment involving medicine, this would be the group that receives the placebo.

Confounding Variables: Also called lurking variables. Other variables that might influence the response variable other than the explanatory variable being studied.

Random assignment: A process for creating similar groups where you take a group of people or objects and randomly split them into two or more groups.

Placebo Effect: The capacity of the human brain to manifest physical responses based on the person believing something is true.

Placebo: A fake medicine or fake treatment used to control the placebo effect.

Chapter 1 Review Problems

1. Tell if the following data is categorical or quantitative and explain why.

    a) The types of cars in the different parking lots.

    b) The average number of hours spent practicing ping-pong.

    c) Areas in North Dakota that have wild mustangs.

    d) Each person is asked if he or she wear glasses, contacts, neither, or both.

    e) The average speed of racecars at the Indianapolis 500.

    f) Exam scores for various students on a history exam.

2. Jim wants to know how much money the average working COC student makes. Describe how Jim could use each of the following techniques to collect data. For each technique, will there be a significant amount of sampling bias or not too much sampling bias?

    a) Systematic

    b) Voluntary Response

    c) Random Sample

    d) Convenience Sample

    e) Cluster Sample

    f) Stratified Sample

    g) Simple Random Sample

    h) Census

3. Define the following key terms and give an example of each.

a) Population

b) Census

c) Sample

d) Random

e) Bias

f) Statistic

4. Describe and give an example of each of the following types of bias. Also state how a person collecting and analyzing data, can avoid these biases.

a) Sampling Bias

b) Question Bias

c) Response Bias

d) Deliberate Bias

e) Non-Response Bias

5. Rachael needs to do an experiment that will show that wearing nicotine patches cause a person to stop smoking. Set up the experiment for Rachael. What is the explanatory variable? What is the response variable? Write a description of the experiment and include the following. What are some confounding variables that she will need to control? How can Rachael control the confounding variables? Include a description of how Rachael use a double blind placebo to control the placebo effect. Describe the treatment group and the control group in the experiment.

6. Compare and contrast the similarities and differences between an experiment and an observational study. How can we tell if we should use an experiment or an observational study?
