Section 1C – Bias

Vocabulary

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

Bias: When data does not reflect the population.

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.
