Sampling and bias

No matter what kind of study we're doing (observational or experimental), we always want to make sure that the subjects in our study are picked randomly if we're using a sample.

Remember that a population N always consists of the entire group of subjects we're interested in. A sample n, on the other hand, is a smaller group within the larger population. So if we're able to study the entire population, then we don't have to worry about the group we look at, because we're looking at everybody, or everything. But if we're only going to be using a sample to represent a larger population, then how we pick the subjects that will be in our sample is very important.

In a perfect world, we want the sample to be representative of the population. If it were in fact perfectly representative, we might call it a **representative sample**, because the information we collect about the sample would "scale up" to the population.

For example, we might want to know how many strawberries are in our fruit salad. The fruit salad fills a 20-cup bowl, so we don't want to pick through the entire thing and count every strawberry. Instead, we'll take a 1-cup scoop, and hope that it's a representative sample of the entire bowl.

In the 1-cup scoop, we count 2 strawberries. If the sample (the cup) perfectly represents the population (the entire bowl), we could "scale up" the strawberry count for 1 cup to 20 cups to get a count of the number of strawberries in the entire bowl.

© Krista King Math

$$\frac{2 \text{ strawberries in our sample}}{1 \text{ cup in our sample}} = \frac{x \text{ strawberries in the bowl}}{20 \text{ cups in the bowl}}$$

$$\frac{2}{1} = \frac{x}{20}$$

$$2 = \frac{x}{20}$$

$$40 = x$$

Therefore, based on the sample, we guess that there are 40 strawberries in the whole bowl. If there are in fact 40 strawberries in the whole bowl, then the single 1-cup scoop would have been a perfectly representative sample of the entire fruit salad.

But this won't always be the case. Maybe the fruit salad wasn't perfectly mixed, and a lot of the strawberries had sunk to the bottom. If we knew that there were actually 100 strawberries in the bowl, then the sample we picked wasn't a very good one.

Many times a sample won't do a good job representing the population because of bias.

Bias in sampling

Bias, by definition, is showing favor toward something over something else. When we talk about **bias** in statistics, we're basically talking about something that skews our results and makes them inaccurate.

© Krista King Math

When we collect data for a sample, and we're using that sample to represent the population, we mentioned before that we want a representative sample. We also call this an **unbiased sample**.

To get a representative, or unbiased, sample, we try to avoid introducing bias into the data. Unfortunately, it's really easy to introduce all different kinds of bias into a data set and skew our results:

Response bias

Measurement bias: There's something wrong with the tool we're using to collect the data, so our method of collecting observations or responses from the sample results in false values. For example, if we calibrated a scale improperly before taking measurements, then all the results would suffer from measurement bias.

Social desirability bias: If our survey asks "Have you ever stolen something?" people may not answer truthfully. If anyone participating in our survey lies when they answer this question, then we have some social desirability bias in our data. This is similar to measurement bias because in both measurement and social desirability bias, there's something wrong with the tool we're using to collect data.

Leading questions: Leading questions are questions that are framed in a way that push respondents toward a particular response. For example, if we ask "Are you more likely to purchase Coca-Cola?" it may cause respondents to answer differently than if we'd simply asked "Are you more likely to purchase another cola brand?" because we're leading them specifically toward Coca-Cola.

© Krista King Math

Undercoverage bias

Selection bias: Also called undercoverage, this is when we don't collect data from an entire group of subjects that should have been included in our data. For example, let's say I own a daycare center and want to find out the mean household income of the families whose children I look after. If I watch 20 children, and I choose to sample only the parents who pick up their kids before $5:00$ p.m., I might be dramatically skewing my data. What if the parents who work later have significantly higher incomes, because anyone who picks up their child before $5:00$ only works part time? I'm not representing all the parents who work late, so that part of the population is under-represented.

Voluntary response sampling: This is when people voluntarily respond to my survey or participate in my study, which means that voluntary response sampling can be a cause of selection bias. People who voluntarily participate may have different habits, tendencies, opinions, or backgrounds than people who tend not to participate. So the data we collect from a sample of voluntary respondents may be biased.

Convenience sampling: This is when we choose a sample simply because it's convenient, not because we're trying to get a good, random representative sample. Therefore, this can be another cause of selection bias. There's almost always some aspect of convenience to sampling, but a good example would be if we're trying to collect data about the people in our city, and we just ask the neighbors who live on our street. It's really convenient to collect

© Krista King Math

data for our street only, but it certainly doesn't give us an unbiased sample for the entire city, so this convenience sample may cause a big problem.

Non-response bias

Non-response bias: This is when we get a large number of people who don't respond to our survey. There may be bias in our data because we didn't collect answers from everyone who didn't respond, and we don't know what they may have said. For example, state representatives often send surveys to all of their constituents to ask them how much they care about different political issues. If they only get a response rate of 5%, that means 95% of constituents didn't bother to send back the survey. Which means the representative only collected opinions from 5% of the population they were interested in, so nonresponse bias may be a big issue.

Direction of bias

Based on the bias that we suspect may exist in our sample (response bias, undercoverage, and non-response bias), we always want to be able to make an educated guess about whether our results are more likely producing an overestimate or an underestimate.

For example, in the response bias section, we talked about the survey question "Have you ever stolen something?" If that question exists in our survey, it could actually lead to response bias or undercoverage, depending on whether people choose to skip the question or lie.

© Krista King Math

Or in the undercoverage section, we talked about surveying only the parents who pick up their children from daycare before $5:00$ p.m. We might admit that we have some undercoverage in our data, and would guess that our estimate for household income is low, since parents who work late might make more money than those who get off work earlier.

Either way, we always want to be thinking about the kind of bias we're introducing, and whether the bias in our data has caused us to overestimate or underestimate the value we're looking at.

Sampling techniques

So how can we avoid bias in our sample data? Well, there are few techniques we can use to divide subjects into groups that help ensure that our data stays as random (unbiased) as possible.

First, we could assign a number to each subject in the population, then pick numbers out of a hat, assigning every other subject to the same group. Or, we could use a random number generator on a computer to do the same thing. We could also get a computer generated string of digits, and then pick out numbers in order from the random number string.

When we assign subjects to groups in a totally random way like this, we call it a **simple random sample**. But even if we assign subjects to groups in a totally random way, we can still end up with a skewed sample. For example, it's possible that we could use a random number generator on a computer to randomly put 50 men and 50 women into two different groups, and yet end up with 40 men and 10 women in one group, and 10 men and 40 women in another group.

© Krista King Math

To fix problems like this, instead of doing a **simple random sample**, we can try to take a **stratified random sample**, where we put some parameter on the sample where we require an even number of subjects from different groups. For example, if we want to have one group of 25 men and one group of 25 women in our sample, then we'll treat men as one population and women as another. Then we'll take a random sample of 25 men from the male population and a random sample of 25 women from the female population. These two groups (men and women) are called the **strata** of the stratified random sample.

Earlier we used the term blocking to describe this. These concepts are the same thing. The strata in sampling are what we called the blocks in an experiment with a randomized block design.

We could also take a **clustered random sample**, where we break our population into clusters, and then either 1) take a random sample within each cluster to be our total sample, or 2) randomly pick some clusters and then sample everyone in those clusters.

In a cluster sample we want each cluster to be similar to the population as a whole. For example, say we have a nicely mixed fruit salad that's been divided into 12 portions. Then each of these portions is a cluster. Once the population is divided into representative clusters you can take a simple random sample of clusters, say 3 out of the 12 portions to analyze the sample.

In the case of a stratified sample, the fruit salad would've had to be separated back into strawberries, bananas, watermelon etc., and a sample from each group selected. So often we will hear a stratified sample is the

© Krista King Math

same *within* groups, (like each fruit is the same) and a cluster sample is the same *between* groups, (like each portion of fruit salad should be representatively the same).

Finally, we could use **systematic sampling**, which is really similar to simple random sampling. The difference is that we assign numbers to individuals in a population and choose them at some specified interval. For example, we could list all students in a school in alphabetical order, choose student #5 as our starting point, and then select every 10th student on the list from there.

© Krista King Math