Sampling Methods and Bias with Surveys: Crash Course Statistics #10

Crash Course: Statistics
https://youtube.com/watch?v=Rf-fIpB4D50
https://nerdfighteria.info/v/Rf-fIpB4D50

===== Introduction (0:0) =====

[Complexly theme] Hi, I'm Adriene Hill. Welcome back to Crash Course Statistics. In our last episode, we talked about how we use experiments to imitate having two parallel universes to test things. But sometimes you can't do certain experiments without becoming an all-powerful and evil dictator.

And since it's statistically unlikely that any of you are evil dictators, today we'll explore those methods. Like we mentioned at the beginning of the series, you're not always able to answer the questions you really want to answer using statistics. For example, it would be great to experimentally test whether getting married increases your lifespan, but you can't randomly assign some people to be marries and force another group to be single.

Not only would that be difficult to enforce, it would also be pretty unethincal–though I suppose you being evil takes care of that particular concern. Similarly, we can't assign someone to be a twin, or a democrat, or a smoker. But that doesn't mean we should just give up and stop trying to find out more about all these topics.

Not at all. Instead, we just need a different method to collect data. Enter non-experimental methods. [Crash Course intro theme]

===== Survey (1:14) =====

One of the most common non-experimental methods is the survey. From user experience surveys on websites to political polls, to health questionnaires at the doctor's office, you've probably taken hundreds of these surveys in your lifetime. There are two things that can make or break a survey: the questions, and who the researcher gives the questions to.

===== Surveys with Unrelated Questions (1:32) =====

The goal of a survey is to get specific information. Say, you're walking your dog in a local park, and someone approaches you and asks you to take a survey on local businesses in your town. When you look at the questions, you notice none of them are about local businesses. Instead, you find yourself answering questions about your politics and religious beliefs.

Unless the surveyor was lying to you about their purposes, this is not a very good survey. It's also not a very good lie. A survey should measure what it claims to measure.

It might seem obvious that having only unrelated questions on your survey is problematic, but there are even more subtle ways a question can be biased.

===== Survey: Multiple Choice Question Example (2:07) =====

= Let's take a look at a few questions from a health survey you might find at the doctor's office. The first question asks you how often you exercise: never, less than 30 minutes a week, or 30 minutes a day. But, what do you answer if you exercise for half an hour, twice a week? Or if you're on swim team and exercise for at least and hour a day?

And does dog walking count as exercise? Multiple choice questions that don't offer all possible options and/or an "other" option can cause respondents to wither skip the question, or feel forced to choose an answer that isn't accurate. Claims made using these questions aren't as strong as they could be if people were offered a full range of choices.

===== Survey: Leading Question Example (2:42) =====

The next question at your doctor's office asks you: "Answer yes or no: I don't smoke because I know it's damaging to my health." This is a leading question since the wording leads towards the quote "desired" answer. This is especially effective when a question deals with sensitive issues like smoking, politics, or religion. People answering the questions want to be seen in a positive light, and so they tend to give the answer they think is "appropriate." While having people fill surveys out anonymously by themselves can help, it can sometimes be the case that respondents don't want to admit things–even to themselves–that are socially undesirable. In general terms, good survey questions are worded in a neutral way, such as asking "How often do you exercise?" or "describe your smoking habits," instead of using wording or options that push survey takers in a certain direction.

And while your doctor wouldn't–or shouldn't–do this, it sometimes groups purposefully use biased questions in their surveys to get the results they want. Apparently, back in 1972, Virginia Slims conducted a poll asking respondents if they would agree with the following statement: "There won't be a woman President of the United States for a long time, and that's probably just as well." Not a well-written question. Biased questions can be more subtle, and can lead to skewed reports of very serious things like sexual assault, or mental health conditions.

It's important to always look for biased questions in surveys, especially when the people giving the survey stand to benefit from a certain response.

===== Sample Population Biases (4:10) =====

Even when researchers have created a non-biased survey, they still need to get it into the right hands. Ideally, a survey should go to a random sample of the population that they're interested in. Usually this means using a random number generator to pick who gets the survey. We do simple random sampling so there's no pattern or system for selecting respondents, and each respondent has an equal chance of being selected.

For example, telephone surveys often use random digital dialing which selects seven random digits and dials them. When someone picks up, they're asked to take a survey.

===== Non-Response Bias (4:38) =====

But here's where we hit our first issue: if people aren't forced to respond to the survey, we might experience something called non-response bias, in which the people who are most likely to complete a survey are systematically different from those who don't. For example, people with non-traditional working schedules like retirees, stay at home parents, or people who work from home might be more likely to answer a middle of the day phone survey. This is a big problem if those groups are different than the population as a whole. If your survey was on health insurance plans or political opinions, it's likely these three groups would have different opinions than the population, but they represent the majority of survey responses, which means your data won't represent the total population all that well.

===== Voluntary Response Bias (5:20) =====

This also related to voluntary response bias, in which people who choose to respond to voluntary surveys they see on Facebook or Twitter are people who again are different than the broad population. This is especially true with things like customer service surveys. People who respond tend to have either very positive or very negative opinions- (I mean, see the comments section below). The majority of customers with an average experience tend not to respond, because service wasn't noteworthy- (wait, does that mean I'm not noteworthy?)

===== Underrepresentation (5:50) =====

Another source of bias is just plain under-representation. If a group of interest is a minority in the population, random sampling paired with response biases might mean that minority isn't represented at all in the sample. Let's say there's a city where five percent of the population is single mothers. It's entirely possible that the sample will contain no single moms.

===== Weighted Responses (6:11)=====

To over come these issues we have a couple options: we could weight people's responses so they match the population, counting the few single mothers who do respond multiple times, so they end up counting for five percent of the total sample. But this can be problematic for the same reason that response bias is problematic. If the few single mothers who respond don't represent all single mothers, our data is still biased.

In a 2016 LA Times / USC political tracking poll, a 19-year-old black man was one of 3,000 panelists, who was interviewed week after week about the upcoming presidential election. Because he was a member of more than one group that was underrepresented in this poll, his response was weighted 30 times more than the average respondent. According to the New York Times, his survey boost his candidates margins by an entire percentage point.

===== Stratified Random Sampling (7:01) =====

Stratified random sampling is another option. It splits the population into groups of interest, and randomly selects people from each of the "stratas" so that each group in the overall sample is represented appropriately. Researchers have used stratified sampling to study differences in the way same-sex and different-sex parents parent their kids. They randomly select people from the same-sex parenting group and randomly select people from the different-sex parenting group to make sure they're well represented in the sample.

===== Cluster Sampling (7:26) =====

Another issue is that getting surveys to people can be expensive. If a cereal company wants to see how families react to their new cereal, it would be costly to send some cereal to a random sample of all the families in the country. Instead, they use cluster sampling which creates clusters -not honey nut clusters- that are naturally occurring -like schools or cities- and randomly select a few clusters to survey instead of randomly selecting individuals. For this to work, clusters cannot be systematically different than the population as a whole, and they should about equally represent all groups.

===== Snowball Sampling (8:00) =====

Issues can also arise when the population being surveyed is very small or difficult to reach, like children with rare genetic disorders or people addicted to certain drugs. In this case surveyors may choose not to use randomness at all, and instead use snowball sampling. That's when current respondents are asked to help recruit people they know from the population of interest, since people tend to know others in their communities, and can help researches get more responses. And note here that these sampling techniques can and are used in experiments as well as surveys.

===== Non-Experimental (8:27) =====

There are other non experimental data collection methods, like a census.

===== Census (8:30) =====

A census is a survey that samples an entire population. The United States conducts a census every ten years, the next one is scheduled to be done in 2020. It attempts to collect data from every single resident of the United States, even undocumented residents and homeless residents. As you can imagine, this is hard, and it's not without error.

In medieval Europe, William I of England conducted a census in order to properly tax the people he had conquered. In fact, a lot of rulers tend to use cenuses(?~8:58) to know just how much money they should be demanding. Until the widespread availability of computers, the U.S. Census data took almost ten years to collect and analyze, meaning that data from the last census wasn't even available until right before the next census. The length of time it took to complete the census is part of the reason we have computers- check out our [Computer Science] series for more on that.

So why collect census data instead of just sampling the population? In the U.S. the census could cost more than 15 billion dollars in 2020. There are a lot of reasons, I mean the Constitution says we have to, but also the census provides the truest measure of the population we can get. It minimizes sampling error. It also functions as a benchmark for future studies. And a census can give researchers really specific information about small groups of the population: information that might be hard to gather with regular sampling methods. And doing statistics on census data is different, because most statistical inference aims to take a small sample, and use it to make guesses about the population But with the census, we already have data from the entire population. We don't need to guess if there are differences, we can just see them. Analysis on census data is usually more concerned with whether differences we see are large enough to make a difference in everyday life, rather than guessing if there is a relationship.

The census, as we said, can take years, and entire countries to fund. That doesn't discount the value of sampling, but we should be cautious. Badly worded polls, fake polls, and biased polls are common- so are the results of those polls. The statistics friendly website FiveThirtyEight put together a great list of advice on how not to fall for a fake poll. Among its advice: Ask yourself if it seems professional. Ccheck to see who conducted the poll and if you trust them. See how the poll was conducted. Check out the questions they asked and who they asked. If it seems fishy, it probably is fishy.

That said, well done surveys are essential. They allow us to get information without all the trouble of doing an experiment. And since they are comparatively easy, they are popular ways for businesses, countries, and even YouTube channels to collect information. In

fact, Crash Course Staticstics has it's own survey, the link is in the description (and here), and it takes way less time than the Nerdfiteria one, I promise.

Thanks for watching, I'll see you next time.

===== Credits (11:07) =====

Crash Course Statistics is filmed in the Chad and Stacey Emigholz Studio in Indianapolis, Indianna, and it is made with the help of all of these nice people. Our animation team is Thought cafe.

If you would like to keep Crash Course free for everyone forever, you can support the series at Patreon: a crowd funding platform that allows you to support the content you love. Thank you to all to our patrons for your continued support.

Crash Course is a production of Complexly. If you like content designed to get you thinking, check out some of our other channels at Complexly.com

Thanks for watching.