Module 2 Video 2: What is a Good Enough Sample?

[00:00:00] Hi, it's Heather and welcome back to another video in our course on how to embed equity and ethics into your data journalism story. And we're in the section about data collection and data sourcing and how to understand the data that you're working with. And one of the things that we're going to talk about today is what the heck is a sample and how do you know if your sample is good enough to put in your data story?

[00:00:31] OK. So what is a sample? A sample is essentially data from a part of the community or group of people that you want to talk about.

[00:00:44] So you want to do a data story on a community. And this community has a lot of people in it, and there some of them are very similar to each other and some of them are very different from each other. And you, as a data journalist, want to use some quantitative information to talk about the attitudes or feelings or experiences of a community. Another word for this community is the population. That can be confusing and technical word. But from a data point of view, the population is the community of people, whether it's a neighborhood or a country. The population is the community that you're interested in talking about.

[00:01:33] Now you as a data journalist, and even if you are a very, very wealthy research institution, you do not have the time, money and capacity to ask every single person in your community and or your population their opinion. So what you do is you take a survey. So you're going to gather some of the attitudes and opinions and experiences from the community and the population. And the way that you take this survey is how you get a sample. It's how you collect a sample.

[00:02:13] So a sample is the group of people who answer your survey or the group of people from within your community, your population that you actually have data from. That's what a sample is. So small group of people or even a large group of people who are a subset of your entire population of interest or your community. The group of people that you have data for. Community sample.

[00:02:42] However, you want to understand the feelings of your community, not the feelings of your sample. So what you're gonna do is you're going to try and use the data or information that you have from your sample to understand your community. And how well this works depends on a couple of things. It depends on how you've collected your sample data. It depends how similar your sample actually is to your community, and it affects how much you can generalize it is also affected by what math you do to the sample data. And when I said whether or not you can generalize it, generalize from the samples to the community. That means whether you can take the information that you have about your sample and say with a high degree of confidence, integrity and ethics, that this sample result actually represents the community or population that you're trying to get it to represent.

[00:03:54] Now the most robust or fancy way to collect a sample is random sampling. If you have truly randomly sampled your data. And your sample size is large enough. Then you can mostly assume that the feelings of your sample reflect the feelings of your community. However, truly random sampling is extremely hard to do. So to truly random sample your community, you need to put everybody in that community into a metaphorical hat and randomly choose people out of that hat. That's very challenging to do. It's very expensive to do. And even if you managed to do it, how are you going to handle the fact that lots of people that you choose don't want to participate in your sample, don't want to give data? So there are a lot of complex reasons that a sample, even if it's a large sample, does not accurately reflect the thoughts and feelings of the community.

[00:05:00] So there's a couple of different things that you need to do with that. And one of them is some fancy math, which is called weighting. So, for example, if your community that you're interested in talking about is 50 50 men and women. But your sample, the group of people that your data about is twenty five percent women and 75 percent men. If there is a difference in the way that men and women are feeling or experiencing, whatever it is that you're collecting data about your sample, which is twenty five women and seventy five men, does not accurately reflect

your community, which is 50 percent men and 50 percent women. So what you need to do is called weighting.

[00:05:48] So you have your community here on the left. They take an online survey. Your sample has more women. I mean, more men than women in it. So you need to do something called weighting your data to make it more accurately reflect the thoughts and feelings of your community. So if you're a data journalist and you're using data that came from a sample, it's very important that, you know, A, how that sample was collected. And then, B, if that sample was weighted and if it was weighted, how was it weighted was weighted according to gender, according to race, immigration status, there's so many different ways that you can weight data. And what this weighting does is it actually changes the results that will be reported. So here on the left, we what we really want to know is how many people in our community think, yes. And how many people in our community think, no. We don't actually know the true answer to that. So we take a survey and we get a small group of people a sample and let's say in the sample there's a lot more yeses than no's. But again, our sample isn't a good reflection yet of our community. We have to weight the data so that it reflects our community. And then you can see that once we've weighted the data, there's a lot more no's and yeses. So the results that you'll get will be very different from a raw sample than a weighted sample. And they'll be very different depending on how you collected the sample, whether or not you collected it randomly or in really any other way.

[00:07:33] This is a quick overview, this is kind of extra credit. You don't really need to know about this, but if you want to know about the different types of sampling, there's random sampling which you've already talked about, which is just pulling names out of a hat. Actually, random. They're stratified sampling where you divide people up into meaningful groups and then pick a few people randomly out of each of those groups. There's volunteers sampling. And that's where you just say, I have a survey. I would really like people to take it. Maybe you put it up online. Maybe it's on your newspaper's Web site. Maybe it's on your own personal Web site. And people who feel like taking it take it. That's volunteers sampling or self selected. This is very, very convenient sampling. And it's ethical because it involves a lot of consent, but it's can be very unrepresentative. The sample that you got through volunteer sampling is not usually generalizable to the whole community. And again, that word generalizable means the information from your sample accurately represents the information from your community. And as a data journalist, that general generalization is something that we do or we're tempted to do quite a lot in writing a story. We have information on a little bit of people. We want to talk about how that affects or is meaningful for a lot of people. And that leap, that generalization is one of the places that you have to be very, very careful about the equity in the ethics. And one of the ways you can check on that is how would your sample collected and was your sample weighted? Those are two ways to just get started on thinking about how a sample can affect your data story's ethics and equity.

[00:09:24] So that's what a sample is, what you need to know about a sample and how to think about a sample in terms of data acquisition. When you're deciding do I want to use this data, can I comfortably use this data in my data story.