
STUDYDADDY
Get Homework Help
From Expert Tutor
Get Help

Project Planner

Data Collection: What Is Sampling?

Title: Data Collection: What Is Sampling?
Originally Published: 2017
Publishing Company: SAGE Publications, Inc.
City: London, United Kingdom
ISBN: 9781526408563
DOI: https://dx.doi.org/10.4135/9781526408563

What Is Sampling?

A sample is a part that stands for the whole. Note that phrase “stands for.” The sample is supposed to represent the whole so that we can say things about the whole on the basis of information about the sample.

Quantitative sampling is always a matter of selecting some cases from all cases. We take a sample from a population. To be able to do that, we need to have some clear idea as to what the population is—the “theoretical universe.” We also need to have some way of accessing (at least) a very large part of that population—the “working universe,” through a list of all cases from which we sample. This is called the **sampling frame.**

In quantitative work we also want the sample to be collected in such a way that we can use statistical inference to attach probability statements to measures derived from the sample and from hypotheses which we test on the basis of sample data.

Why is it important to have a random sample?

We want to be able to make statements about the population from which the sample is drawn on the basis of measurements from the sample. That is, we want to make statements about parameters (population values) on the basis of statistics (sample-based estimates of population values). To do this, we must draw our sample randomly, or in a way which is very close to random.

Why does the size of the sample matter?

A random sample of size N is a sample drawn in such a way that all samples of size N have an equal chance of being drawn. It is not a sample drawn in such a way that all cases have an equal chance of being drawn.

Our sample size must be large enough (N greater than 40). Then, regardless of the shape of the distribution of any parameter in the population, if we were to draw all possible samples of that size from the population, the estimates of the parameter value would be normally distributed. It would have a mean equal to the parameter value and a standard deviation equal to the standard deviation of the population divided by the square root of N-1. This is a consequence of the law of large numbers. It is the basis on which we can make probability statements about the population.

What happens if the sample size is too small?

If the sample size N is too small (if N is less than 40), we can only make probability statements if we can assume that the parameter of interest is normally distributed in the population itself. This requires us to use special tests. For example, the T-test tests for differences of means when we have two small unmatched samples. These tests are less powerful, and less able to detect differences in values of a parameter in the population for different groups. Note that, for larger samples, we have to increase the size of a sample by four times to double the power, increasing the ability to detect statistically significant differences.

Assigning numbers to cases

To conduct a fully random sample, we should assign a number to all cases on our sampling frame. Then we use a random numbers table or generator to give us numbers for N cases. Systematic sampling is a process where, instead of doing this, we first work out the fraction of the population which we need to sample to get our sample of size N. So if we want to draw a sample of 1,000 from a population of 100,000, then we are sampling one hundredth of the cases.

In systematic sampling we would then randomly select a number between 1 and 100, say we get 57. So we start with the 57th case and then take every 100th case going forward—157, 257, and so on. Systematic sampling ensures that our sample is of cases spread through the population as that population is ordered on the sampling frame. Intuitively, this seems more appropriate than the entirely possible random sample of the first 1,000 cases only. So long as the frame is not organized in a way which might bias the sample, the systematic sampling can be considered as equivalent to random sampling and for geographically ordered frames, for example lists of addresses, it does spread the sample out across the whole population.

Can we tell if a sample is representative?

A random sample is more likely to be representative of the population from which it is drawn. If we think of all possible samples we could draw, then there are far more which are close to the character of the population than are far from it. However, when we have a sample, all the information we have about the population comes from that sample. So we have no way of knowing if the sample is representative or not.

How can we achieve representativeness?

We try to achieve representativeness by using information in the sampling frame as a basis of organizing the selection of cases. We try to make that selection representative in terms of the variation in the population for which we have information from the frame.

For example, if we are drawing a sample of students in a University and our sampling frame is drawn from registration information, then we are likely to know:

- Gender
- Subject of study
- Age
- Year of study
- Whether undergraduate, taught postgraduate or research student.

We can sort our cases into categories on the basis of these five dimensions. For example, we might look at female politics students aged less than 20 in their second year of study who are undergraduates. From the sampling frame we will know what proportion of all students fall into that set. We then sample within the category. This process is called stratified sampling because we have sorted the population into categories.

Proportionate and disproportionate sampling

We can sample proportionately so if that category included 1 in 50 of all students, then we would take one 50th of our sample from it. However, we can also sample disproportionately, taking larger numbers from small and interesting categories and smaller numbers from large categories. Then, in order to give an account of the characteristics of the population, we must weight any data from the category so that they represent the correct proportion of cases. So if, instead of selecting a number of female politics students which would produce one 50th of our sample, we selected double that number, then we would have weight results from this category by 0.5.

As long as we select randomly within our categories, we can treat the results from a stratified sample as results from a random sample. Stratification actually increases the power of a sample but generally most researchers treat them as equivalent to a simple random sample.

Often we don't have any useful information about our individual cases in the sampling frame, but we do have information about some category to which they belong. For example, if sampling households from a list of addresses, we know where they are located depending on their state or local authority ward. Typically, census data will give us a lot of aggregate information about the set of households in the ward. For example, we can find out the proportion of households owning or renting, and if they are renting from a social landlord. We can sort the wards into categories by tenure profile, draw a stratified sample of wards according to this classification, and then sample for households within the wards thus selected. This would be a two-stage stratified sample.

Birds of a feather flock together. We must be careful when we sample entities containing other entities of interest to us. The most common example of this is when we have people living in households. If we take a sample of households and then make statements on the basis of information about the characteristics of the relevant population of households, then we are on firm ground. If we collect information from all individuals in the households and then make statements on the basis of that information about the population of individuals, then we are in error. People in large households will be over-represented in the sample compared with people in small households. We have sampled clusters of individuals, not individuals. There is a correction formula which gets round this issue and guides sampling from multi-case entities—be aware of the issue and use it if necessary.

Quota sampling is a technique used by commercial research and, in particular, marketing research. It is intended to generate representative samples. Information about the population from which the sample is to be drawn is used to sort that population into categories. The number of cases to be sampled in each category is then determined. This is the quota. Cases are sought to achieve that number by asking screening questions of potential informants. When the quota is full then sampling for that category stops.

Quota samples are not random samples. Therefore, it is impossible to use the tools of statistical inference to attach probabilities to statements made on the basis of the data they generate. For this reason, they are very seldom used in scientific research.

When doing qualitative research, we are generally interested in getting results which are in some way representative. So, if we are going to interview a set of informants using qualitative procedures, we generally try to think about the whole population of potential interviewees and make our selection in a way which gives us a representative set.

The same holds for any qualitative technique intended to enable us to generalize beyond the actual cases for which we have information. But here the idea of sampling can be extended beyond our normal understanding of cases—pre-existing real entities such as people, schools, firms, etc.—to "instances" or examples of social action which are of interest to us.

This approach is characteristic of grounded theory where the objective is to generate a theoretical sample by saturation. Saturation means that we keep looking for new instances of interest until we stop finding anything different. So our sample is representative not because of our initial selection, but because of the way we work when we are actually doing the research. The process is different but the objective is the same—we want to get a part which can stand for the whole.

This term is used in a specific sense in grounded theory, as developed by Glaser and Strauss (1967). The object here is not to achieve representativeness but rather to develop theoretical accounts of social reality, often causal accounts, by seeking instances which are different than those already investigated.

Search for resources about sampling