The relationship between sample and population

Earlier, you learned that **inferential statistics** use sample data to draw conclusions or make predictions about a larger population. Data professionals use inferential statistics to gain valuable insights about their data.

In this reading, you'll learn about the relationship between sample and population in more detail. We'll also discuss how data professionals use sampling in data work, and the importance of working with a sample that is representative of the population.

Population and Sample

Population vs. sample

In statistics, a **population** includes every possible element that you are interested in measuring, or the entire dataset that you want to draw conclusions about. A statistical population can refer to any type of data, including:

- People
- Organizations
- Objects
- Events
- And more

For instance, a population might be the set of:

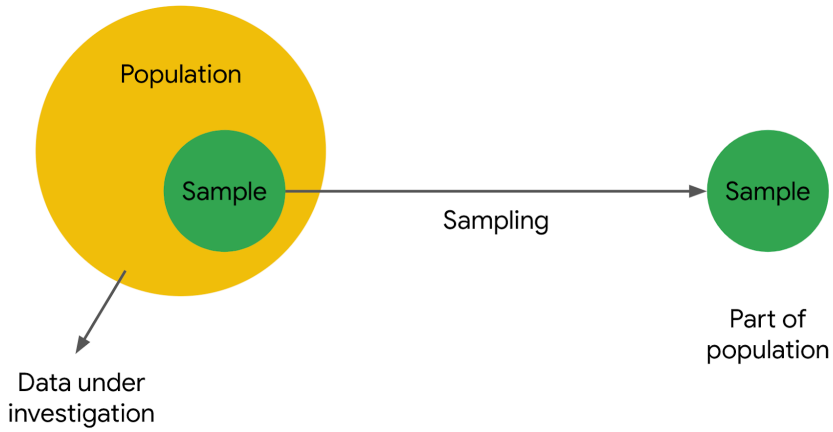
- All students at a university
- All the cell phones ever manufactured by a company
- All the forests on Earth

A **sample** is a subset of a population.

Samples drawn from the above populations might be:

- The math majors at the university
- The cell phones manufactured by the company in the last week
- The forests in Canada

Data professionals use samples to make inferences about populations. In other words, they use the data they collect from a small part of the population to draw conclusions about the population as a whole.



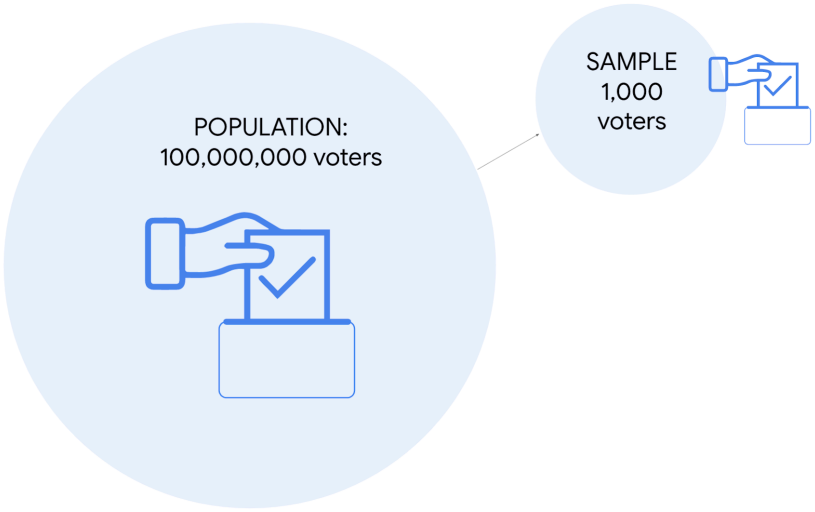
Sampling

Sampling is the process of selecting a subset of data from a population.
In practice, it's often difficult to collect data on every member or element of an entire population. A population may be very large, geographically spread out, or otherwise difficult to access. Instead, you can use sample data to draw conclusions, make estimates, or test hypotheses about the population as a whole.
Data professionals use sampling because:

- It's often impossible or impractical to collect data on the whole population due to size, complexity, or lack of accessibility
- It's easier, faster, and more efficient to collect data from a sample
- Using a sample saves money and resources
- Storing, organizing, and analyzing smaller datasets is usually easier, faster, and more reliable than dealing with extremely large datasets

Example: election poll

Imagine you're a data professional working in a country with a large population like India, Indonesia, the United States, or Brazil. There is an upcoming national election for president. You want to conduct an election poll to see which candidate voters prefer. Let's say the population of eligible voters is 100 million people. To survey 100 million people on their voting preferences would take an enormous amount of time, money, and resources – even assuming it would be possible to locate and contact all voters, and that all voters would be willing to participate. However, it is realistic to survey a sample of 100 or 1000 voters drawn from the larger population of all voters. When you're dealing with a large population, sampling can help you make valid inferences about the population as a whole.



Representative sample

To make valid inferences or accurate predictions about a population, your sample should be representative of the population as a whole. Recall that a **representative sample** accurately reflects the characteristics of a population. The inferences and predictions you make about your population are based on your sample data. If your sample doesn't accurately reflect your population, then your inferences will not be reliable, and your predictions will not be accurate. And this can lead to negative outcomes for stakeholders and organizations.
Statistical methods such as probability sampling help ensure your sample is representative by collecting random samples from the various groups within a population. These methods help reduce sampling bias and increase the validity of your results. You'll learn more about sampling methods later on.

Example: election poll

Ideally, the sample for your election poll will accurately reflect the characteristics of the overall voter population. A voter population in a large country will be diverse in political perspectives, geographic location, age, gender, race, education level, socioeconomic status, etc. Your sample will not be representative if you only collect data from people who belong to certain groups and not others. For example, if you survey people from one political party, or who have advanced degrees, or are older than 70. The results of an election poll based on a non-representative sample will not be accurate. In general, any claims or inferences you make about any population will have more validity if they are based on a representative sample.