Samples and Populations

When we analyze data, we usually assume that values in a data set are from a sample drawn from a larger population.

A population is the complete set of observations, or the entire group of objects, that you are interested in understanding. You are ultimately interested in understanding the population, but you typically can't measure every member of the population.

Instead, you select a sample of observations from the population and use this sample to estimate characteristics of the population.

These population characteristics are called parameters. Parameters are usually denoted with Greek letters, like μ and σ.

In the previous lesson, you learned about measures of center and spread based on sample data.

These measures are called statistics.

Statistics are usually denoted by letters from the English alphabet.

The average of the population is called the population mean. To distinguish it from the sample mean, it is denoted by the Greek letter μ.

The sample mean x-bar is calculated from sample data and is an estimate of the population mean.

The sample standard deviation (s) is an estimate of the population standard deviation, which is denoted by the Greek letter σ.

Similarly, the sample variance (s-squared) is an estimate of the population variance, denoted by σ-squared.

In this course, and in statistics texts, if you see a Greek symbol, you know that we are talking about the unknown, or theoretical, population characteristic.

If you see the English letter, you know that we're talking about a value of the characteristic that was calculated using sample data.

So, for example, if you see the symbol μ you know that this is a theoretical value, and if you see x-bar, you know that this is a value that is calculated using data.

Let's talk a little more about samples.

Because you're using sample statistics to estimate population parameters, the sample should be representative of the population.

That is, the characteristics of the sample should be similar to the characteristics of the population.

Using sample data that are not representative of the population leads you to draw incorrect conclusions about the population characteristics.

Simple random sampling can help ensure that the sample is representative of the population.

With simple random sampling, every member of the population has an equal chance of being selected for the sample.

A result of simple random sampling is that the sample statistics are good estimates of the population parameters. You learn what "good" means in an upcoming video.

There is a tendency to collect data that are the most readily available or the easiest to collect.

This is called convenience sampling, and the resulting samples might not be representative of the population.

For example, say you're studying dimensions of metal parts.

Batches of parts are transported in bins.

With convenience sampling, you might select a sample of parts from the top of the bin and measure the dimensions of these parts.

What's wrong with this approach?

The parts at the top might be the last parts off the production line. As a result, this sample might not represent the entire batch of parts.

Your estimates of the dimensions of parts in the batch might be biased, and your conclusions about the quality of the parts in the batch might be incorrect.

Another sampling technique is to systematically select a sample. For example, you could choose every 10th item off a production line to measure. Systematic sampling can produce representative samples and is often more practical than simple random sampling.

The most important thing to keep in mind when collecting data for analysis is to make sure that your sample data are representative of the population that you are studying. Otherwise, your conclusions about the population characteristics might not be valid.

Statistical Thinking for Industrial Problem Solving

Close