Introduction to Statistical Inference

Recall the idea of a population, which was introduced in the Exploratory Data Analysis module. A population is a collection of objects, individuals, events, or items that interest us.

Consider the Impurity scenario, involving the production of a polymer. A catalyst is required for the chemical reactions to occur to produce the polymer. The catalyst contains a chemical that can create an impurity in the polymer. The lower the % impurity, the better, because lower impurity equals higher yield.

An improvement team is tasked with decreasing the average impurity to 3% and ensuring that all product meets the upper specification limit of 7%.

In this example, the population is the batches of polymer that this process produces. At any point in time, only a given number of batches have been produced. Assuming that the process hasn't changed, you analyze the impurity for these batches in order to draw conclusions, or make inferences, about the overall process output.

In many cases, the population is conceptual, such as all of the parts that might be produced, or it's extremely large. This can make it too costly or otherwise impossible to measure every member of the population. In these cases, you use a subset from the population for data analysis, and you draw inferences about the population from this subset.

As an analogy, think about a blood test. The doctor wants to analyze your blood chemistry. You can't send all of your blood to the lab for testing. Instead, the doctor extracts a relatively small sample of blood, and analyzes this sample to draw conclusions about your overall blood chemistry.

However, if you can measure every member of the population, then you don't need to use samples. You can simply calculate the important characteristics using all of the available data.

For example, at the end of a baseball or football season, you have the data for the entire season. Because you know all of the population characteristics, you don't have to infer these characteristics using samples.

In this module, you learn basic methods for drawing inferences from sample data.

The methods of inference essentially break down into two basic activities: interval estimation and hypothesis testing. There are many variations of these methods, and you select the method that is appropriate for the question you are asking and the types of data you have.

If you're asking the question, "what is the unknown value?", then you would compute an interval estimate of the value.

If you're asking the question, "is the unknown value greater than X?", then you would conduct a hypothesis test.

In the next video, you are introduced to perhaps the most commonly used type of interval estimate, a confidence interval. We use a familiar example involving categorical data: political polling results.

However, for the remainder of this lesson, we focus on basic methods involving continuous, numeric data.

For a more comprehensive discussion of inferential methods for both continuous and categorical data, see the Read About It for this module.

Statistical Thinking for Industrial Problem Solving

Close