Visualizing Sampling Variation

Interval estimation is rooted in the simple idea that random samples vary from one another.

This variability produces sampling error. Sampling error is the imprecision arising from using samples rather than the entire population. To illustrate, we simulate random sampling from a population using the Sampling Distribution of Sample Means script.

This script is in the JMP Help menu, under Sample Data, then Teaching Scripts, and then Interactive Teaching Modules.

Recall that the file Impurity.jmp has data on the impurity of 100 batches of a polymer. The mean is 6.1232 and the standard deviation is 1.5142.

For this simulation, we'll treat the Impurity data as if it were a population, with a mean of 6.1232 and a standard deviation of 1.5142. We'll simulate many samples with 100 observations and study the variation in the means for these samples.

First, we enter these values for the population mean and the population standard deviation. The Impurity distribution is slightly right skewed, so we'll also change the shape of the distribution to right skewed, and change the name of the variable to Impurity.

Then we change the sample size to 100, which is the number of observations in the Impurity data set, and click Draw Additional Samples to draw one sample with 100 observations.

The population data are highly right skewed. So, as expected, the first sample is right skewed. Notice that the mean of this sample is not precisely 6.1232, although it is pretty close.

As we draw additional samples, we see that the means of the individual samples are all close to the population mean, and they are distributed around the population mean.

Let's change the number of samples to 2000 to draw many samples at one time. After thousands of repetitions, the distribution of all sample means, also called the sampling distribution of the mean, approaches the familiar symmetric mound shape, and is centered at the known population mean.

To better see this sampling distribution, we save the sample means to a data table and create a distribution of these means.

Notice that the average of the sample means is nearly identical to the mean of the population. This distribution is approximately normal, even though the underlying population is very right skewed.

You can also see that approximately 95% of the sample means fall between 5.8 and 6.4. This is the idea behind a confidence interval.

If we were to repeatedly sample from this population, approximately 95% of the sample means would fall within the interval of 5.8 to 6.4, and approximately 5% of the sample means would fall outside of this interval.

Let's tie this back to the impurity scenario. In reality, the Impurity data are just one sample, a sample of 100 observations, rather than a population. From this one sample, we estimate the true mean impurity using a point estimate and a confidence interval.

You can see that the 95% confidence interval for the mean is nearly identical to the range of values in which 95% of our simulated sample means fell.

This confidence interval has captured the uncertainty in our estimate caused by sampling variation.

Statistical Thinking for Industrial Problem Solving

Close