The Central Limit Theorem

The Central Limit Theorem makes the normal distribution particularly useful for statistical analyses.

When you analyze data, you are analyzing one sample of data from a population.

If you had selected a different sample, your computed sample statistics would be slightly different.

The Central Limit Theorem enables you to understand the behavior of different random samples drawn from the same population.

Suppose you have a variable, X, that has a distribution with population mean μ and population standard deviation σ. You draw many samples of size n from this distribution.

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.

The Central Limit Theorem tells you that the distribution of the means of these samples becomes more normal as the sample size increases.

This distribution of sample means is centered at the true mean, μ, and has a population standard deviation, σ over the square root of n, that becomes smaller as the sample size increases.

To illustrate the Central Limit Theorem, we use a simulation. Here, we've simulated data from a population that is extremely right-skewed.

The mean of the population is 0.2 and the standard deviation is 0.267.

This is a histogram of 1000 individual values drawn from this population. The mean is 0.197, and the standard deviation is 0.265. These values are close to the true population values, and the distribution is very right-skewed.

Now, we select samples of size 5 from this same population. This is a distribution of 1000 sample means, where each mean is calculated from five observations. That is, each of the 1000 observations in the histogram is a sample mean.

The first thing you notice is that the distribution looks far less skewed than the distribution of individual values. It is more mounded in shape and has a much shorter tail. The distribution of sample means is close to the population mean, and the standard deviation of the sample means, 0.124, is much smaller than the standard deviation of individual values.

Here are the distributions of sample means where the sample size is 10, 50, and 100. Uniform scaling has been applied to make it easier to compare these distributions.

Notice that the means of these distributions are very close to the population mean of 0.2, that these distributions become more normal as the sample size increases, and that the spread or variability in the sample means becomes smaller as the sample size increases.

We estimate the standard deviation of sample means by dividing the sample standard deviation by the square root of the sample size. This value is called the standard error of the mean, or the standard error. It tells us that the spread of the distribution of sample means becomes smaller by a

factor of the square root of n as the sample size increases.

This explains why we often use sample averages in statistics. We get more precise estimates of population behavior by taking samples. Larger samples give more precise estimates than smaller samples. The standard deviation of the sample mean is always less than the standard deviation of the raw data (by a factor of the square root of the sample size). And sample averages are approximately normally distributed, even if the underlying distribution is not normal, and as the sample size increases, this distribution becomes more normal. Averaging is also an effective noise filter, enabling you to focus in and see the signal through the noise of the variability in the raw data.

You'll see evidence of the Central Limit Theorem at work in in the Quality Methods module, and again when you learn about statistical intervals and hypothesis tests in the Decision Making with Data module.

In the next practice exercise, you explore the Central Limit Theorem using the Sampling Distribution of Sample Means teaching module.

Statistical Thinking for Industrial Problem Solving

Close