Central Limit Theorem

The central limit theorem states that the sampling distribution of the mean of any [independent](http://stattrek.com/help/glossary.aspx?target=independent) [random variable](http://stattrek.com/help/glossary.aspx?target=random_variable) will be normal or nearly normal, if the sample size is large enough.

How large is "large enough"? The answer depends on two factors.

* Requirements for accuracy. The more closely the sampling distribution needs to resemble a normal distribution, the more sample points will be required.
* The shape of the underlying population. The more closely the original population resembles a normal distribution, the fewer sample points will be required.

In practice, some statisticians say that a sample size of 30 is large enough when the population distribution is roughly bell-shaped. Others recommend a sample size of at least 40. But if the original population is distinctly not normal (e.g., is badly skewed, has multiple peaks, and/or has outliers), researchers like the sample size to be even larger.

Confidence Interval

A *confidence interval* gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data. The common notation for the parameter in question is http://www.stat.yale.edu/Courses/1997-98/101/theta.gif. Often, this parameter is the population mean http://www.stat.yale.edu/Courses/1997-98/101/mu2.gif, which is estimated through the sample mean http://www.stat.yale.edu/Courses/1997-98/101/xbar.gif. The *level C* of a confidence interval gives the probability that the interval produced by the method employed includes the true value of the parameter http://www.stat.yale.edu/Courses/1997-98/101/theta.gif.

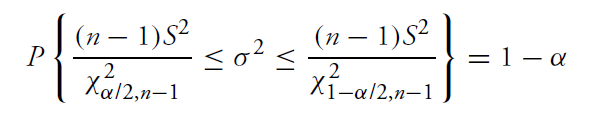
## Confidence Intervals for Variances ~~and Standard Deviations~~

We have learned that estimates of population means can be made from sample means, and confidence intervals can be constructed to better describe those estimates. Similarly, we can estimate a population standard deviation from a sample standard deviation, and when the original population is normally distributed, we can construct confidence intervals of the standard deviation as well.

### The Theory

Variances and standard deviations are a very different type of measure than an average, so we can expect some major differences in the way estimates are made. We know that the population variance formula, when used on a sample, does not give an unbiased estimate of the population variance. In fact, it tends to underestimate the actual population variance. For that reason, there are two formulas for variance, one for a population and one for a sample. The sample variance formula is an unbiased estimator of the population variance. (Unfortunately, the sample standard deviation is still a biased estimator.)

Also, both variance and standard deviation are nonnegative numbers. Since neither can take on a negative value, the domain of the probability distribution for either one is not   (−∞,∞),   thus the normal distribution cannot be the distribution of a variance or a standard deviation. The correct PDF must have a domain of   [0,∞).   It can be shown that if the original population of data is normally distributed, then the expression   (n−1)s2/σ2   has a chi-square distribution with   n−1   degrees of freedom.



It is worth noting that since the chi-square distribution is not symmetric, we will be obtaining confidence intervals that are not symmetric about the point estimate.

## What Is Goodness-of-Fit for a Linear Model?

Definition: Residual = Observed value - Fitted value

Linear regression calculates an equation that minimizes the distance between the fitted line and all of the data points. Technically, ordinary least squares (OLS) regression minimizes the sum of the squared residuals.

In general, a model fits the data well if the differences between the observed values and the model's predicted values are small and unbiased.

Before you look at the statistical measures for goodness-of-fit, you should [check the residual plots](http://blog.minitab.com/blog/adventures-in-statistics/why-you-need-to-check-your-residual-plots-for-regression-analysis). Residual plots can reveal unwanted residual patterns that indicate biased results more effectively than numbers. When your residual plots pass muster, you can trust your numerical results and check the goodness-of-fit statistics.

## What Is R-squared?

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model. Or:

R-squared = Explained variation / Total variation

R-squared is always between 0 and 100%:

* 0% indicates that the model explains none of the variability of the response data around its mean.
* 100% indicates that the model explains all the variability of the response data around its mean.

In general, the higher the R-squared, the better the model fits your data. However, there are important conditions for this guideline that I’ll talk about both in this post and my next post.