Imagine that we have the following regression equation: .

In the equation above, *y* is the dependent variable, *x* is the predictor, is the population y-intercept, is the population slope parameter, and is the residual term – that is, the term which allows the population regression line to fall above or below actual data points. Said differently, is the difference between observed values of *y* and the values of *y* predicted by the regression model (denoted by ).

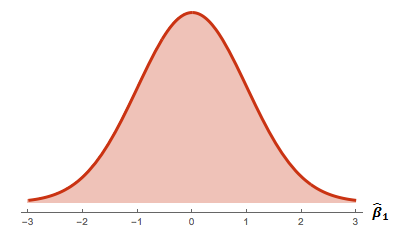
We first estimate and by minimizing the sum of squared residuals, .

Recall that and are quantities calculated from sample data, and are point estimators of the true corresponding population parameters, and .

After we estimate and using our sample data, we can do hypothesis tests. For example, below is a test about the parameter:

vs.

Because is a quantity calculated from sample data, if we were to have an infinitely large number of samples drawn from the population when the is true (i.e., where , we would have a bunch of different values of (i.e., each sample would have its own value). That is, we could think of as a random variable whose values vary from sample to sample, and whose values have a distribution of their own. If is true and , the distribution of might look like this:



Notice that the distribution looks normal, and is centered around 0. That is, – i.e., the expected value, or *true population average*, of is (which, under is 0).

We know that when is true (i.e., when , the quantity

has a t-distribution with *n – 2* degrees of freedom, where *n* is the sample size.

So, we can calculate the t-statistic using the formula above, and then obtain the p-value, which is the probability of obtaining the that is as different from 0 (value stated in ) as the one we obtain in our sample.

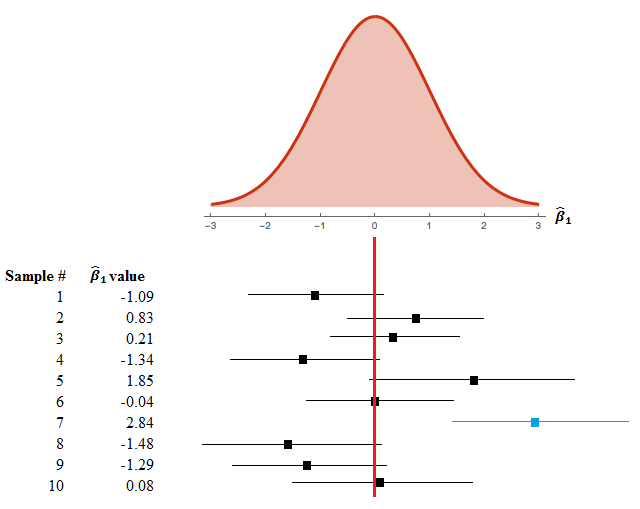
We can also calculate the 95% confidence interval for using the following formula:

For our intents and purposes, the calculation of the t-statistic, the p-value, and the confidence intervals will be done in R.

However, it is worth noting that the reason statisticians like to have confidence intervals is that they can have a range of plausible values for our population parameter , instead of just a single estimate, .

The 95% confidence interval may be interpreted as follows: if we were to have an infinitely large number of samples, each with its own value of , and we were to construct 95% confidence intervals for from each sample, then 95% of the resulting confidence intervals would contain the true population parameter and the remaining 5% would not. Since we are dealing with a single sample, we do not know whether our sample is part of the 95% of the samples that contain the true value of , or part of the 5% of the samples that don’t, but if it’s part of the 5%, we can simply consider ourselves unlucky.

In the figure below, we have a bunch of samples drawn from the population where (i.e., when the is true). Each little square below the histogram of represents the value of from each sample. The “whiskers” represent the 95% confidence interval for each sample. If we were to have a very large number of samples (and not just 10) then 95% of all the confidence intervals would not contain , the true population value under .



Another thing worth noting: if 0 is part of the 95% confidence interval, then we will *not* be able to reject at the level α = 0.05 significance level.

On the other hand, if 0 is *not* part of the 95% confidence interval, then we will be able to reject for at the level α = 0.05 significance level.