Confidence Intervals

Definition

A C% confidence interval for a population parameter is an interval of numbers such that, if we could generate many different random samples from the population, C% of the samples would contain the true value of the population parameter. A confidence interval provides an interval estimate for a population parameter. To calculate a confidence interval, you must specify the level of confidence, you must have a random sample from the population, and you must know the sampling distribution of the statistic that forms the basis for the confidence interval.

Theoretical example

We are interested in estimating the mean of a population, µ. The population has a standard deviation of σ. We have a random sample of N=100 observations from this population. The sample mean, , has a normal distribution with mean µ and standard deviation . The 95% confidence interval for µ is . We obtain the number 1.96 from a table giving values for a standard normal distribution, using the fact that the probability that a standard normal random variable lies in the interval [-1.96, 1.96] is 0.95.

The idea is that provides a good estimate of the population parameter µ. A confidence interval provides another estimator of µ that takes into account the variability of . Once we have calculated a particular confidence interval given an observed sample, we are 95% confident that our interval contains the true value of the population parameter.

In general, for statistics that have a normal distribution, the general formula for a confidence interval is [Statistic - Margin of Error, Statistic + Margin of Error], where Margin of Error = (standard error of the statistic) x (table value). For most cases, we use a t-table to find the appropriate table value for a specified level of significance.

Practical example

We are interested in estimating the average amount, µ, spent by teenage shoppers at an online music store in a one-month period with 90% confidence. We know that the standard deviation of purchases is $16. We collect a random sample of purchases of 64 teens. The mean value of purchases of these 64 teens is a random variable which has a normal distribution with a mean of µ and standard deviation of $16/8 = $2. In one particular sample, the 64 teens purchased an average of $56 worth of music online. The $56 is the sample average for this particular sample. A different sample of the purchases of 64 teens would have produced a different sample average.

A 90% confidence interval for µ is given by [ $56 - 1.645 x $2, $56 + 1.645 x $2 ] = [ $52.71, $58.29 ]. The number 1.645 is obtained from a standard normal table: the probability that a standard normal random variable is in the interval [-1.645, 1.645] is 90%. We are 90% confident that the population average purchases by teen shoppers is between $52.71 and $58.29.

The best estimate of the population mean is the sample mean, $56. The confidence interval allows us to consider the amount of uncertainty that we have about this estimate. Although our best estimate of µ is $56, we are 90% confident that the value of µ is in the interval [$52.71, $58.29]. This is an estimate that takes into account our uncertainty about our estimation.

General example

In many cases, the formula for a confidence interval is (statistic - TV*SE, statistic + TV*SE), where TV is shorthand for table value and SE is shorthand for the Standard Error of the Statistic. The table value is determined by the sampling distribution of the statistic you are using.

Properties

- As our level of confidence increases, the width of the interval increases and the estimate becomes less precise. Specifically, a 90% confidence interval is wider than an 80% confidence interval.
- If we keep the level of confidence the same and if the standard deviation stays the same, an increase in the sample size, N, reduces the width of the confidence interval and the estimate becomes more precise.
- The sample statistic is always contained in the interval. In fact, the sample statistic is in the middle of the confidence interval.
- If we create many different samples, and calculate a confidence interval for each sample, then C% of those samples will contain the true population parameter. 100 - C% of those intervals will NOT contain the true population parameter value.

Why can’t I say that there is a 95% probability that the interval contains the population parameter?

This is actually a good question, subject to some debate among statisticians. In most elementary statistics classes, the population parameter is a fixed (but unknown) number. It's not a random variable. Since it's not a random variable, you can't assign probabilities to the values it takes except in a vacuous sense. By “vacuous” sense, I mean the following: suppose the true value of the population mean is 12.4. Then the probability that = 12.4 is 1; the probability that equals any other value is 0.
In other words, a confidence interval either contains the population parameter (and the probability that it is in the interval is 1.0) or it does not contain the value (and the probability that it is in the interval is 0).

The way to think about confidence intervals is to think about creating many different intervals from many different samples. If we have a 95% CI, then 95% of the intervals calculated from the set of samples will contain the true population parameter and about 5% will not. If we pick ONE sample and calculate ONE interval, we are 95% confident that we got one of the samples that does contain the population parameter.

Many students seem to want to say that a wider interval is more accurate. I think that what they mean is that an interval that has a higher confidence level (and is, thus, wider) is more likely to contain the population parameter. This is another version of interpreting a confidence interval with probability ideas, and is incorrect.