

Confidence Intervals Explained

By: Perceptive Sciences Corporation

If you've collected data from a sample of users and computed summary statistics (e.g., average task time, average error rate, etc.), what you really want to know is how well the values computed from this sample reflect the rates for the entire population of users. The solution is to compute confidence intervals. For example, a 90% confidence interval might tell you that the mean time for a given task - in the general population - falls somewhere between 2 & 3 minutes. Doubling the sample size may allow you to be 90% certain that the population value falls between 2.3 & 2.7 minutes. Note that by increasing the sample size, you shrink the range of possible values (i.e., increase the precision of the estimate). You can also compute intervals for other confidence levels (80%, 95%, or whatever you like - just not 100%). You can select the sample size that gives you the confidence level and precision you would like.

The question: How well do my statistics predict population values?

If you've collected some sample user data, you can calculate statistics to characterize the results (e.g. mean time on task, mean error rate, etc.). But what you really want to know is how well these statistics reflect the rates in the general population of users. You're not interested in comparing these values to any other values, you simply want to know how good each statistic is on its own as a population estimator. The reason the sample mean is insufficient as an estimate is because another sample might yield a slightly different estimate of task time (i.e., have a different mean), and a third sample may yield yet another slightly different mean.

The number of people in our sample is directly related to the credibility of our estimates. The more people we have in our sample the better. However, there are limits (e.g., time, money) to the number of people we can include in our test. What we need is a measure of the reliability of our statistics given different sample sizes so that we can make choices about the number of people to include (i.e., how big a sample do we need to be satisfied?).

The solution: Compute confidence intervals.

By computing confidence intervals, you can conclude that the real population value is somewhere between two values (the upper and lower confidence bounds), and you can specify the degree of confidence you'd like to have (e.g., 80%, 90%, 99% confidence - but never 100%). For example, I could tell you that I'm 90% sure that tomorrow's temperature will be between 85 & 95 degrees, or I could tell you that I'm 95% sure that the temperature will be between 80 & 100 degrees. Note that in the second case, the confidence is higher, but the estimate is less precise. Determining which estimate is better is a matter of opinion.

Here's another example. If you have a sample of 10 users, and mean time is 2.5 minutes for a given task, computing a 90% confidence interval might lead to the following conclusion: I can be 90% sure that the mean time for the entire population of users would be somewhere between 2 & 3 minutes. A 95% confidence interval might say the following: I can be 95% certain that the true population value is between 1.5 & 3.5 minutes.

What you really want is to have high confidence that the population value falls somewhere in the interval estimate, and for the interval to be as precise (small) as possible. The way to get both of these is to increase the sample size - the larger the sample, the more precise the confidence intervals will be. For example, if the sample size is doubled, the 90% confidence interval might shrink to 2.3 & 2.7, and the 95% confidence interval might shrink to 1.9 & 3.1.

Technical note: Where do these things come from?

It would be nice if we could simply make a table showing the intervals to expect for different confidence levels and sample sizes.

Unfortunately, it's not that easy. The actual confidence interval values are a function of the variability in the users' performance, and so some information regarding the population variance must be gotten from, you guessed it, a sample of data!!! This could be based on results from a previous study (of the product in question), or else from a new, relatively small sample. From there, the decision could be made as to whether the resulting confidence is adequate, or if additional users should be tested to improve the estimate (i.e. to shrink the range of values likely to contain the true population mean).

Perceptive Sciences Corporation is a science based market research, user interface design, and user testing firm, employing experts in the fields of cognitive psychology, information sciences and human factors studies. Perceptive Sciences serves best-in-class technology based companies and market leaders in a wide range of industries in the U.S. and Europe.