How to Calculate a Confidence Interval

Dean Wheeler
Brigham Young University
July 2019

Introduction

A confidence interval is a measure of how certain we are that if we repeated a set of measurements, we would get an equivalent result. Confidence interval can either be expressed as a numerical range or as *best estimate* $\pm$ *margin of error*. The sample mean (algebraic mean of the repeated measurements) is usually taken as the best estimate.

The margin of error accounts for uncertainty due to imperfect sampling and is a measure of precision. When you measure a quantity repeatedly, you tend to get different results, i.e. variance. This could be due to natural fluctuations in the process, random environmental noise, intrinsic limitations in an electronic sensor, or human reading of an analog instrument. When you collect multiple data points and then *average* them, this reduces the susceptibility to such random variations and improves precision, which is why is it always a good idea to collect as much data as you can.

Accuracy requires both *precision* and *trueness*. Using an average from many data points does not eliminate systematic errors or bias, for instance if a temperature sensor were mis-calibrated and always produced values that were too high. Here I describe how to characterize random fluctuations; characterizing systematic error is a topic for another time.

Independent Samples

To generate a mean with a low margin of error, you must collect many independent samples for a single measured quantity. An independent sample means that its residual ($x - \bar{x}$) is not correlated with other sample residuals, because you repeated the experiment with *enough time or distance* or other intervening process separating the samples so that they are each subject to a different set of environmental disturbances. For instance, if you used a computer to collect 100 temperature readings from a single temperature sensor over the course of 100 seconds, then adjacent readings would not be independent from each other if whatever is causing temperature fluctuations in the environment takes longer than 1 second to occur.

One way to improve independence and reduce systematic error is to design your experiment so that variables you control are arranged in random order. In other words, instead of collecting pressure data for multiple flow rates arranged from smallest to largest, you should put the flow rates in random order. Or if you are trying to make multiple measurements at the same flow rate, then you could move the control valve randomly higher or lower between measurements, so that they are more independent and do not suffer from hysteresis (a type of bias caused by a valve always being driven in the same direction).

To test for independence, one can plot residuals ($x - \bar{x}$) for samples to see if any undesired

patterns emerge. For a time series (i.e. samples collected with a fixed time interval), a Durbin-Watson test can be done to detect unwanted autocorrelation or serial correlation.

Margin of Error

Following accepted engineering practice, you need to calculate a confidence interval in the form

$$\bar{x} \pm \left(t \frac{s}{\sqrt{n}} \right) \tag{1}$$

where $\bar{x}$ is the sample mean, t is the critical value from Student's t distribution, n is the number of independent samples, and s is the standard deviation of the samples.

The quantity t can be looked up on a table or from an Excel function (see below) and depends on the *confidence level* and on a *degrees of freedom* value, which is $df = n - 1$. You should generally use a confidence level of 95% or higher and use a "two-tailed" distribution. 95% confidence level is equivalent to a *level of significance* of $\alpha = 0.05$; you will know you calculated t correctly if you get $t \approx 2$ for $df > 27$.

To use Excel to calculate quantities needed in Eq. 1, use the following functions

Quantity	Excel Function
$\bar{x}$	AVERAGE(*range*)
s	STDEV.S(*range*)
n	COUNT(*range*)
t	T.INV.2T(α, df)

where *range* indicates the range of cells containing the sample measurements, such as A2:A20.

Notice in Eq. 1 that the margin of error (the quantity after the $\pm$) decreases with increasing value of n. This reflects the fact that you can have more confidence in $\bar{x}$ if it is based on a larger number of independent samples. However, if your samples are not fully independent (e.g. there is correlation between the sample residuals) it's as if n is incorrectly too large and Eq. 1 will underestimate the true margin of error.

Level of Precision

When presenting your confidence interval, you should round off your margin of error to 1 or 2 significant digits. One rule of thumb is if the leading digit in the margin of error is a smaller number (1 or 2) then use two significant digits and otherwise use one significant digit. In any case, you should then round off your sample mean so its precision matches the precision of your margin of error. For example, $\bar{x} = 131.773 \pm 2.4329$ becomes 131.8 ± 2.4. Notice how the revised sample mean and margin of error don't have the same number of significant digits, but they do have the same level of precision (i.e. decimal place of least significant digit) and that the sample mean was appropriately rounded off to that level of precision.

2-Sample T-Test

There are instances where you take a set of measurements, make some change to the process, and then take a second set of measurements. The question is whether the process change led to a *statistically significant change* in the respective means from the two data sets. This is known as a 2-sample or independent sample t-test. In its most general case, we are comparing two different data sets and we want to know the difference between the means and the corresponding margin of error for this difference:

$$(\bar{x}_2 - \bar{x}_1) \pm \left(t \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right) \tag{2}$$

where $\bar{x}_k$, s_k, and n_k are respectively the mean, standard deviation, and number of samples for data set k. The t statistic is computed as we did above, typically with $\alpha = 0.05$ and an assumption of a "two-tailed" distribution. The degrees of freedom value for computing t is given by

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{s_1^4}{n_1^2(n_1 - 1)} + \frac{s_2^4}{n_2^2(n_2 - 1)}} \tag{3}$$

This number is not necessarily an integer: to get a t value one can either interpolate on a t-table or simply use the Excel function T.INV.2T, which does not require an integer.

Once a confidence interval is computed by Eq. 2, this can be compared to a hypothesized difference (called the null hypothesis), typically zero difference. For instance, if Eq. 2 produced the result 8 ± 15 at the 95% confidence level, then this means that a difference of zero is within the interval and we conclude that the process did not have a statistically significant effect, or that any observed change was due to random chance. On the other hand, if the result of Eq. 2 were 8 ± 3, then a change of zero is outside of the confidence interval and we can conclude that the process did have a statistically significant effect.

Eqs. 2 and 3 are known as Welch's t-test and are considered fairly robust as long as df is not too small. The following alternative t-test assumes the two sets of measurements have the same population standard deviation (i.e. s_1 and s_2 would converge to the same value if enough samples are taken). In that case, one can use a "pooled" value of the standard deviation, s_p, for the confidence interval:

$$(\bar{x}_2 - \bar{x}_1) \pm \left(t \, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \tag{4}$$

where

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2} \tag{5}$$

and the t statistic uses $df = n_1 + n_2 - 2$.