Confidence intervals

Thomas H. Meyer

8/25/2025

Probability foundations

This article discusses confidence intervals of a random variable. Random variables draw their values from some underlying probability distribution, and, in this note, we'll use the normal probability distribution. The normal probability distribution, or just "normal distribution" for short, is characterized by its **mean** (average) and its **variance**. The mean is the distribution's most probable value. As values are observed from a random variable drawing its values from a normal distribution, those values will cluster around the mean. In theory, a normally distributed random variable can produce *any* value from positive infinity to negative infinity; however, although possible, values farther and farther away from the mean become more and more unlikely the farther away they are. The distribution's variance describes the spread of the values.

The totality of every possible outcome that a distribution can produce is called its population. A population is a theoretical concept because it is an infinite set. It's impossible to observe every value in an infinite set, so we must draw samples from the population and draw inferences about the distribution from what we see in our samples.

Confidence intervals

Surveyors and engineers work to tolerances. We know that we cannot set a stake *exactly* at the coordinates stipulated in the design, and a hole in a girder cannot be drilled at *exactly* some location on the girder. The hole and the stake are considered to be in the right location if they are within some tolerance, plus or minus, of the design location. Confidence intervals are like this. For example, suppose a surveyor uses an EDM to observe slope distances between the same two stations many times with a different setup each time. Every observation will yield a different value for the slope distance. None of the observations will be the true value and, even if one of them—by some miracle—did happen to be the true value, we would never know it. The mean (average) of the observations is the best estimate of the true distance but, especially if there are very few observations, it is possible that the true distance falls outside the set of observations. Were this to happen, the mean would not be a good estimate of the true value: the set of observations must bracket the true value for the mean of the set to be representative of the true value. A confidence interval gives us the spread of values needed to be confident at a given level-of-confidence that the interval contains the true mean.

The normal distribution is ubiquitous in the statistical treatment of land-surveying measurements; however, since a normally distributed random variable can produce any value, this assumption is usually not entirely appropriate. For example, we model the observations of an electronic distance meter (EDM) with a normally distributed random variable, but a properly functioning EDM cannot produce wild readings that are very far from the true distance, much less negative readings. Properly functioning surveying instruments produce readings that tend to cluster rather closely to the true value, so the variability of the readings is small. Such readings would be modeled by a distribution whose variance is quite small compared to the mean, so it's extremely unlikely to observe a negative number from a random variable drawing from such a distribution. As such, the normal distribution assumption usually works quite well, but there are exceptions. Hereafter, we use the normal distribution exclusively.

There are random variables that produce values that are not **independent**, which means the prior values affect subsequent values. The discussion in this note assumes that all observations *are* independent.

The mean

Suppose we have a sample of *n* independent observations taken from a normally distributed random variable x. This could be, say, $n = 10$ observations of slope distance given by an EDM, where the instrument and the target were releveled for each observation to ensure independence of observations. Let's denote this as a mathematical set that we can write as $X = \{x_1, x_2, \dots, x_n\}$ where each x_i is an outcome from observing *x*. For the example, each x_i would be an observed slope distance. The **mean** of *X* is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

The mean is the best possible estimate of the true average of the probability distribution that *x* is drawn from given the information available, which is the sample set *X*. For surveyors, we could say that the mean is the best estimate of the quantity we are trying to measure, like a slope distance. The Central Limit Theorem establishes that, for a "large" sample size, the sample mean approaches the population mean with absolute certainty, with a probability equal to exactly 1. Here, "large" means that the sample mean approaches the population mean with absolute certainty as *n* becomes "infinite." Since the sample mean is guaranteed to approach the true value, we can rely on the mean to be a "good" estimate if the sample set is "large enough". It also means that, since we cannot make an infinite number of measurements, we cannot know the true slope distance. We can only make estimates of the true value.

Since we know that the mean is not the true value, we would like to know how far the mean is from the true value. This, too, is not knowable exactly without an infinite number of observations. However, like the mean, this variability can also be estimated from the sample set *X*. We can estimate the amount of variability in *x* by computing the **variance** of the sample. The variance of a set of random samples *X* is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}(x_1 - \bar{x} + x_2 - \bar{x} + \cdots + x_n - \bar{x})$$

and the **standard deviation** of *X* is $s = \sqrt{s^2}$. The $1/(n-1)$ term is called Bessel's correction, and it adjusts the estimate of the variance to account for sample variance's tendency to underestimate the population variance when the mean is also estimated from the same data.

A confidence interval gives us the spread of values needed to be confident at a given level-of-confidence that the interval contains the true mean.

Having assumed that our data are normally distributed with a certain mean and variance that we estimated from the sample, we can now use mathematics to determine how likely it is that observations fall within some certain range of that distribution. The normal distribution's probability density function (pdf) is described as a "bell curve" because it is tall in the middle and flattens out way from the middle in the "tails." The area under that curve between some upper and lower limits is the probability of observing a value from that distribution between the upper and lower limits. Conversely, we can ask what limits are needed to encompass some percentage of area of the density function. We can set a probability, called a **confidence level**, say, 95%, and use mathematics to determine the values that bound the area of the density function equal to the confidence level. As the level of confidence increases, the span widens but we are more confident that the interval spanned by the area contains the true mean. This interval is called a **confidence interval**. The formula for the confidence interval of a sample of n independent observations taken from a normally distributed random variable x is

$$CI = \bar{x} \pm t\frac{s}{\sqrt{n}}$$

where

CI denotes two values: two because of the $\pm$ in the formula. They are the higher and lower limits of the confidence interval,

$\bar{x}$ is the average (mean) of the samples

t is a number called a **critical value** that is associated with the confidence level value that we choose. In practice, we would usually get t's value from a table of **degrees of freedom** (to be defined later) vs. confidence level from Student's t distribution. More about this below but, for now, t's value gets larger and larger as we ask for higher and higher confidence levels. If the confidence level is 1, meaning that we insist that the confidence interval is *guaranteed* to include the true mean, then t will become "infinitely" large. This is in keeping with the notion that we need an infinite number of samples to know the true mean.

s is the standard deviation of the sample

n is the number of samples

Here are some things to notice about the formula.

- This is not the formula for standard deviation, so, obviously, confidence intervals are not just scaled standard deviations.
- Confidence intervals include a confidence level, which standard deviations do not.
- The confidence interval is centered around the mean, as it should be, because the mean is the best estimate of the true value.
- Smaller standard deviations give tighter confidence-interval limits. This makes sense. If the data have a small amount of variability, then the data suggests we have better knowledge of the true value.
- CI becomes tighter as n increases. The more samples we take, the more we know about the distribution.
- CI gets wider as t increases. This means that, if we want to be more and more confident that the confidence interval contains the true mean, then we have to make the interval wider.

Twice the standard deviation, and Student's *t* distribution

We find the t-values in a critical-value table for Student's t distribution, or one can use statistical software packages that have this functionality built in. Such a table gives t values for various degrees of freedom at various confidence levels. The number of degrees of freedom equals $n - 1$.

It was stated above that t's value comes from Student's t distribution, but we've been assuming that x is normally distributed. What's going on? Student's t distribution is almost identical to the normal distribution, but its probability density function's shape is slightly different from the equivalent (same mean and variance) normal distribution's pdf to account for using a sample to estimate the distribution's mean and variance instead of using the true—but unknowable—values. We could, in practice, use a critical-value table for a normal distribution, and the difference would be quite small. Even though the difference is quite small, we use Student's *t* distribution to be as formally correct as possible, especially because the process of finding *t*'s value is no more complicated using Student's *t* than using a normal distribution.

Here are some values for *t* at various confidence levels and sample sizes. At a 95% confidence level, we see that the *t* values are around 2.

		Sample Size (n)				
		10	15	20	30	100
confidence	90%	1.833	1.761	1.729	1.699	1.66
	95%	2.262	2.145	2.093	2.045	1.984
	99%	3.25	2.977	2.861	2.756	2.626

Summary

This is where the notion comes from that a confidence interval is simply twice the standard deviation because the formula for *CI* contains a $t \times s$ term and, at a confidence level of 95%, $t \approx 2$. However, the term is actually $t\ s/\sqrt{n}$. Here, *n* is shrinking the confidence interval to reflect the fact that having more samples leads to better knowledge (a tighter interval) than having fewer samples. This is *not* accounted for in the variance formula despite the presence of a $1/(n-1)$ term multiplying the summation. This term, the Bessel correction, is there to adjust for a statistical bias, and it has nothing to do with more samples producing more reliable estimates. But there is more to this.

Land surveyors usually don't take a large number of observations as they work, so *n* is often less than five. Suppose $n = 4$, then $\sqrt{n} = \sqrt{4} = 2$, and dividing the sample standard deviation by 2 moves the upper and lower bounds of the confidence interval closer together by half the standard deviation. In this case, the width of the confidence interval equals the twice standard deviation (twice because $t \approx 2$), so the rule-of-thumb that a confidence interval is just twice standard deviation is very nearly true. There are scenarios when this would not be a good rule-of-thumb, such as observing a check mark many times over the course of a job, or using a global navigation satellite system receiver to collect two hours' worth of data at 30-second epochs. As always, the surveyor should understand the theory behind the statistics they use to ensure that short cuts do no harm.