

Confidence interval for a population mean

We've learned how to find the sample mean and sample proportion, and we understand that these are sample statistics that we can use to estimate corresponding population parameters.

But as we mentioned before, a sample mean or sample proportion might be a great estimate of the population parameter, or it might be a really bad estimate. So it would be really helpful to be able to say how confident we are about how well the sample statistic is estimating the population parameter. That's where confidence intervals come in.

Point and interval estimates

The sample mean and sample proportion are both examples of a **point estimate**, because they estimate a particular point. The benefit of using a point estimate is that it's easy to calculate. The drawback is that calculating a point estimate doesn't give you any idea of how good the estimate really is. The point estimate could be a really good estimate or a really bad estimate, and we wouldn't know it either way.

In contrast, we can find an **interval estimate**, which instead gives us a range of values in which the population parameter may lie. It's a little harder to calculate than a point estimate, but it gives us much more information. With an interval estimate, we're able to make statements like "I'm 95 % confident that the population mean lies in the interval (a, b) ," or "I'm 99 % confident that the population proportion lies in the interval (a, b) ."

These 95 % and 99 % values we're referring to are called confidence levels. A **confidence level** is the probability that an interval estimate will include the population parameter. It's most common to choose 90 % , 95 % , or 99 % as your confidence level, and then find the interval that's associated with that particular confidence level.

It's important to clarify what we mean when we talk about a particular confidence level. To use an example, if we choose a 95 % confidence level, what we're saying is that 95 % of all confidence intervals that we find will contain the population parameter. Of if we choose a 99 % confidence level, we're saying that 99 % of the confidence intervals we build will contain the population parameter.

The alpha value α

So for a 95 % confidence level, we're saying that 5 % of the confidence intervals we construct won't contain the population parameter. This 5 % (or 10 % for a 90 % confidence level, or 1 % for a 99 % confidence level) is called the **alpha value**, α . We also call it the **level of significance**, or the probability of making a Type I error (more on Type I and Type II errors later). So

$$\alpha = 1 - \text{confidence level}$$

Or put another way, a $1 - \alpha$ confidence interval has a significance level of α .

The confidence interval when σ is known

When population standard deviation σ is known, the **confidence interval** is given as (a, b) by

$$(a, b) = \bar{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}}$$

where (a, b) is the confidence interval, \bar{x} is the sample mean, z^* is the **critical value** (which is the z -score for the confidence level you've chosen), σ is population standard deviation, and n is your sample size.

You'll also see this formula written as

$$(a, b) = \bar{x} \pm z^* \sigma_{\bar{x}}$$

since sample standard deviation is $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. You'll even see the same formula written as

$$(a, b) = \bar{x} \pm z_{\alpha/2} \sigma_{\bar{x}}$$

No matter how you write the formula, the confidence interval is given by the sample mean \bar{x} , plus or minus the **margin of error**, so the margin of error is

$$z^* \cdot \frac{\sigma}{\sqrt{n}} = z^* \sigma_{\bar{x}} = z_{\alpha/2} \sigma_{\bar{x}}$$

If we examine the confidence interval formula, we see that the confidence interval is related to the confidence level (as given by z^*), the standard error σ , and the sample size n .

From the formula we can see that:

- The higher the confidence level, the wider the confidence interval.
- The larger the standard error σ , the wider the confidence interval.
- The larger the sample size n , the smaller the standard error σ , and therefore the narrower the confidence interval.

In general, we want the smallest confidence interval we can get, because the smaller the confidence interval, the more accurately we can estimate the population parameter.

Region of rejection

When we talked about margin of error, we said that it could be written as $z_{\alpha/2}\sigma_{\bar{x}}$. The α refers to the alpha value α that we talked about earlier.

When we pick, for example, a 95 % confidence level, we know that the alpha value is $1 - 95 \% = 5 \%$. If we have an alpha value of 5 %, that means we can expect the smallest 2.5 % and largest 2.5 % of values to fall outside the confidence interval. Because $\alpha = 0.05$, that means $\alpha/2 = 0.05/2 = 0.025$ of the area under the far left of the probability distribution, and $\alpha/2 = 0.05/2 = 0.025$ of the area under the far right of the probability distribution, will fall outside the 95 % confidence interval.

Using a z -table, the z -values associated with +0.025 and -0.025 are +1.96 and -1.96 . Which means the boundaries of the 95 % confidence interval are $z_{\alpha/2} = 1.96$ and $-z_{\alpha/2} = -1.96$.

From this, we can conclude that any z -value outside of $z = \pm 1.96$ will put us outside the 95 % confidence interval, and inside the **region of rejection**. So $\pm z_{\alpha/2}$ are the boundaries of the region of rejection.

Since we'll use them all the time, it's a good idea to know the z -values that will give us the boundaries of the region of rejection for these common confidence levels.

For a 99 % confidence level, $z = \pm 2.58$

For a 95 % confidence level, $z = \pm 1.96$

For a 90 % confidence level, $z = \pm 1.65$

Example

A machine is filling water bottles, and the amount of water in the bottles is normally distributed with a standard deviation of $\sigma = 1$ ounce. You take a sample of 100 bottles and find that the bottles are filled with an average of 16 ounces. What is the confidence interval for a confidence level of 90 % ?

We're asking the amount of water in ounces that correspond to an upper and lower limit for an area of 90 % in the center of the normal distribution. Which means the confidence interval will leave out 5 % of the area under the distribution in the left tail, and 5 % of the area under the distribution in the right tail.

If we look up z -scores that correspond to 5 % on the lower end, and 95 % on the upper end, we get $z = \pm 1.65$.

If we then plug everything we know into the confidence interval formula, we get

$$(a, b) = \bar{x} \pm z^* \cdot \frac{\sigma}{\sqrt{n}}$$

$$(a, b) = 16 \pm 1.65 \cdot \frac{1}{\sqrt{100}}$$

$$(a, b) = 16 \pm 1.65 \cdot \frac{1}{10}$$

$$(a, b) = 16 \pm 1.65(0.1)$$

$$(a, b) = 16 \pm 0.165$$

Therefore, we can say that the confidence interval is

$$(a, b) = (15.835, 16.165)$$

$$(a, b) \approx (15.8, 16.2)$$

We could also express this as the sample mean plus or minus the margin of error, or 16 ± 0.2 ounces. We're 90% certain that the actual population mean of water in the bottles is between 15.8 and 16.2 ounces.

Required sample size for fixed margin of error

Often we'll want to determine the smallest possible sample we can take in order to stick to a specific margin of error. We can easily find the sample

size by manipulating the margin of error formula and then plugging in a few values. The margin of error formula is

$$ME = z^* \cdot \frac{\sigma}{\sqrt{n}}$$

Since we want to find a sample size, solve this for n (which represents sample size).

$$ME\sqrt{n} = z^*\sigma$$

$$\sqrt{n} = \frac{z^*\sigma}{ME}$$

$$n = \left(\frac{z^*\sigma}{ME} \right)^2$$

Now, let's say for instance that we're solving a problem where we want a 95 % confidence interval (corresponding to a z -score of 1.96), let the standard deviation be 5.14, and we want a margin of error of ± 2 , then the smallest possible sample size we can take to ensure that margin of error is

$$n = \left(\frac{1.96 \cdot 5.14}{2} \right)^2$$

$$n = 5.0372^2$$

$$n \approx 25.37$$

To meet that threshold, and keep a margin of error of ± 2 at 95 % confidence, we'd need to take a sample size of at least $n = 26$.

When σ is unknown and/or you have a small sample

If you need to find a confidence interval for a population mean, but you don't know the population standard deviation σ , then you can no longer use the normal distribution and a z -table to find a critical value for your confidence interval.

Instead, you'll need to use the t -distribution. The t -distribution is similar to the normal distribution, in the sense that it's bell-shaped and symmetrical around the mean, and the area under the curve is 1.0.

But the exact shape of the t -distribution depends on the number of **degrees of freedom** (df), which is given by $n - 1$, where n is the sample size. With a small number of degrees of freedom, the t -distribution is flatter than the normal distribution. Once the degrees of freedom reaches about 30, the t -distribution follows the normal distribution almost exactly, which is why we use 30 as the cutoff value between a “small” sample size and a “large” sample size.

Under this set of circumstances, the confidence interval will be given by

$$(a, b) = \bar{x} \pm t^* \cdot \frac{s}{\sqrt{n}}$$

We can find a critical value for t by cross referencing the number of degrees of freedom with the α value or the confidence level in what we call the **student's t -distribution**, which is similar to the z -table.

We'll only use the t -distribution when population standard deviation is unknown and therefore must be approximated by sample standard deviation s , and/or when the sample size is less than 30.