Confidence interval for the 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 the values of their associated 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 point estimate for the population mean, μ, is the sample mean, $\bar{x}$, and the point estimate for population standard deviation, σ, is sample standard deviation, s.

The benefit of using a point estimate is that it's easy to calculate. The drawback is that calculating a point estimate doesn't tell us how good or bad the estimate really is. The point estimate could be a really good estimate or a really bad estimate, and we wouldn't know one way or the other.

In contrast, we can find an **interval estimate**, which 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 the confidence level, and then find the interval associated with that 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.

Alpha α and the region of rejection

To take the inverse of the last statement, for a $95\,\%$ confidence level, we're saying that $5\,\%$ of the confidence intervals we find 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

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

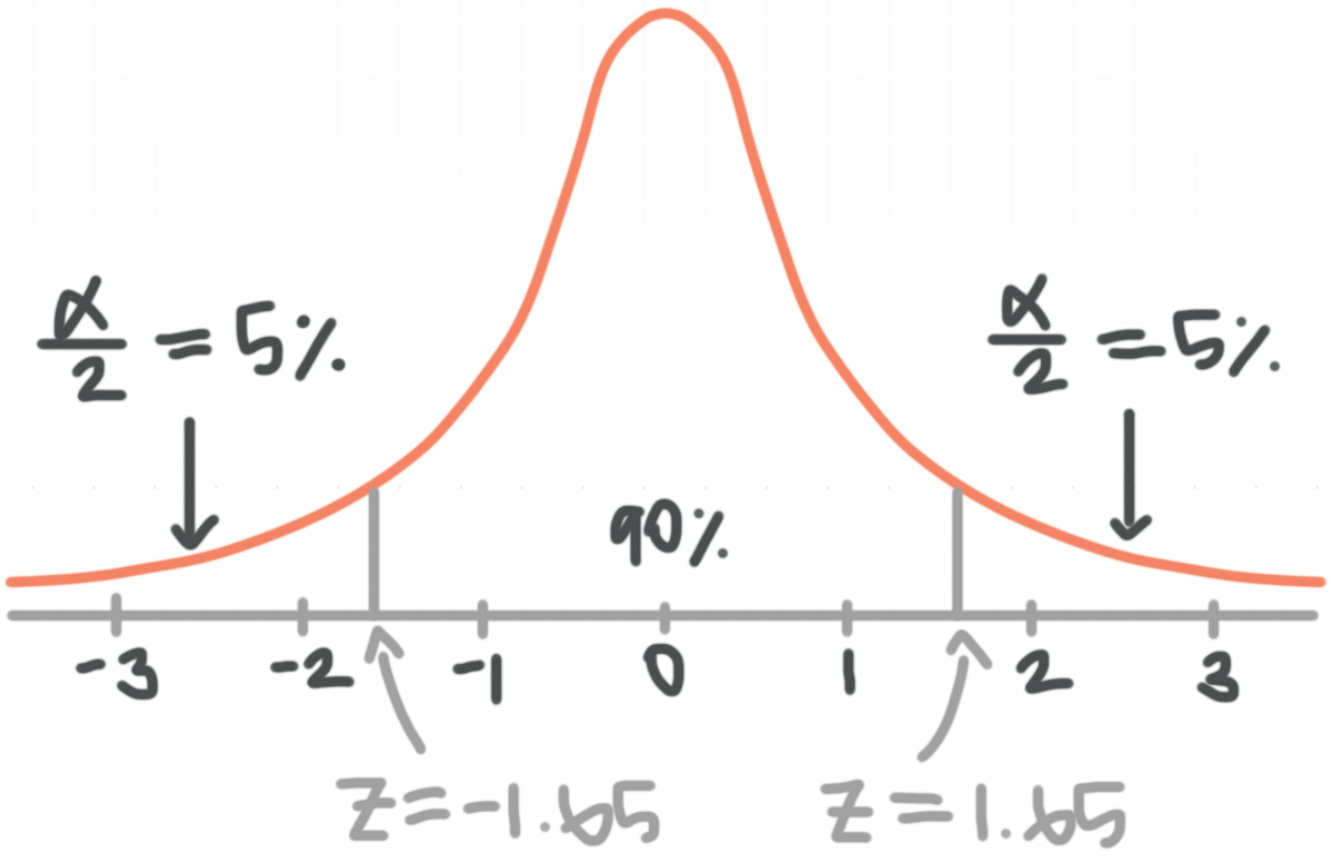
Put another way, a $1 - \alpha$ confidence interval has a significance level of α.

We can visualize α as the total area under the normal distribution outside of the confidence interval. For instance, given a $90\,\%$ confidence level, the alpha value is

$$\alpha = 1 - 0.90$$

$$\alpha = 0.10$$

Since the confidence interval is always centered around the mean of the normal distribution, we can show the central $90\,\%$ of the distribution, with half of α in the lower tail to the left of the confidence interval, and the other half of α in the upper tail to the right of the confidence interval.



In other words, at a $90\,\%$ confidence level, we can expect the smallest $5\,\%$ and largest $5\,\%$ of values to fall outside the confidence interval, because α is split evenly into the upper and lower tails, and $\alpha/2 = 0.10/2 = 0.05$.

Using a z-table, the z-values associated with -0.05 and $+0.05$ are -1.65 and $+1.65$, respectively. Which means the boundaries of the $90\,\%$ confidence interval are $-z_{\alpha/2} = -1.65$ and $z_{\alpha/2} = +1.65$.

From this, we can conclude that any z-value outside of $z = \pm 1.65$ will put us outside the $90\,\%$ 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 $90\,\%$ confidence level, $z = \pm 1.65$

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

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

The confidence interval when σ is known

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

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

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 we've chosen), σ is population standard deviation, and n is the sample size. Since the

standard deviation of the sampling distribution of the sample mean (standard error) is $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, we'll also see this formula written as

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

In both of these versions of the formula for the confidence interval, we'll sometimes use $z_{\alpha/2}$ instead of z^*. They mean the same thing, so using one versus the other isn't actually changing the formula. Using the $z_{\alpha/2}$ notation is an easy way to remember that the α value gets cut in half, with half of α in the lower tail and half of α in the upper tail, to form the region of rejection.

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

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

Now that we know the confidence interval formula, what's the formula actually telling us? Well, if we examine the confidence interval formula, we see that the confidence interval is related to the confidence level (as given by z^*), the population standard deviation σ, and the sample size n.

From the formula, we can see that:

- The higher the confidence level, the wider the confidence interval (because as z^* gets larger, the margin of error will get larger, which makes the entire confidence interval wider).

- The larger the population standard deviation σ, the wider the confidence interval (because as σ gets larger, the margin of error will get larger, which makes the entire confidence interval wider).
- The larger the sample size n, the narrower the confidence interval (because as n gets larger, the margin of error will get smaller, which makes the entire confidence interval narrower).

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.

Keep in mind that the finite population correction factor applies to the confidence interval formula in the same way that it applied to the formula for standard error.

If sampling is done without replacement from a finite population, then the confidence interval formula we use is

$$(a, b) = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

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

When σ is unknown, we use the best available approximation, which is sample standard deviation, s. But because s is a less reliable predictor of σ than σ itself, we have to use a more conservative t-value, instead of a z -value, to find the confidence interval.

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

Similarly, if our sample size is small ($n < 30$), then we don't have enough data for the Central Limit Theorem to reliably apply, and we'll again have to use the more conservative t-value, instead of a z-value, in our confidence interval formula. So our confidence interval formula is

$$(a, b) = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} \qquad \text{whenever } \sigma \text{ is known}$$

$$(a, b) = \bar{x} \pm t^* \frac{s}{\sqrt{n}} \qquad \text{whenever } \sigma \text{ is unknown, and/or } n < 30$$

Let's do an example where we find the confidence interval around the mean when population standard deviation is unknown, and the sample size is small.

Example

The mean exam score of a sample of 10 randomly selected students is 86.7, with a sample standard deviation of 5.72. Determine the confidence interval of the true mean at a confidence level of $99\,\%$.

Because population standard deviation is unknown, and our sample size is small, we'll have to use the confidence interval formula with a t-score instead of with a z-score.

We're given the sample mean and the sample standard deviation. The only thing we need to find is the t-value, which depends on the degrees of freedom and the confidence level. In our case,

$$\text{df} = n - 1 = 10 - 1 = 9$$

Since the confidence level is $99\,\%$, the confidence interval will leave out $0.5\,\%$ of the area under the t-distribution in the left tail, and $0.5\,\%$ of the area under the t-distribution in the right tail.

Look up the critical t-value in the t-table.

	Upper-tail probability p									
df	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	**3.250**	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence level C									

We see that $t = 3.250$. Substitute the values we've found into the formula for the confidence interval.

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

$$(a, b) = 86.7 \pm 3.250 \cdot \frac{5.72}{\sqrt{10}}$$

$$(a, b) \approx 86.7 \pm 5.88$$

Therefore, we can say that the confidence interval is

$$(a, b) \approx (86.7 - 5.88, 86.7 + 5.88)$$

$$(a, b) \approx (80.82, 92.58)$$

We're $99\,\%$ certain that the mean exam score for the population falls between 80.82 and 92.58.

Let's do an example in which population standard deviation σ is known.

Example

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

Because population standard deviation is known, we can use the confidence interval formula with a z-score.

We're asking for 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 $z = \pm 1.65$. Now plug everything we know into the confidence interval formula.

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

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

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

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

Therefore, we can say that the confidence interval is

$$(a, b) = (16 - 0.165, 16 + 0.165)$$

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

We're $90\,\%$ certain that the actual population mean of the amount of water in the bottles is between 15.835 and 16.165 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

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

Since we want to find a sample size, we'll solve this for n.

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

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

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

Now let's say, for example, that we're solving a problem where we want a $95\,\%$ confidence interval (corresponding to a z-score of 1.96), that the standard deviation is 5.14, and that 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 $n = 26$. Keep in mind that if we assume the population is normally distributed or that we're sampling with replacement, then $n = 26$ is acceptable. But if the population is not normal and we're sampling without replacement, then it would be

safer to round up to $n = 30$ to satisfy the CLT condition and ensure the validity of using the normal approximation.