Construct a confidence interval for a small sample size

So far, you’ve constructed confidence intervals for large sample sizes, which are usually defined as sample sizes of 30 or more items. For example, when you estimated the mean battery life of a new cell phone, you used a random sample of 100 phones. On the other hand, small sample sizes are usually defined as having fewer than 30 items. Typically, data professionals try to work with large sample sizes because they give more precise estimates. But, it’s not always possible to work with a large sample. In practice, collecting data is often expensive and time-consuming. If you don’t have the time, money, or resources to take a large sample, you may end up working with a small sample. In this reading, you’ll learn how to construct a confidence interval for a small sample size. We’ll go step-by-step through an example involving mean emission levels for a new car engine.

Large versus small sample sizes

First, let’s briefly discuss the different methods you use to construct confidence intervals for large and small sample sizes.

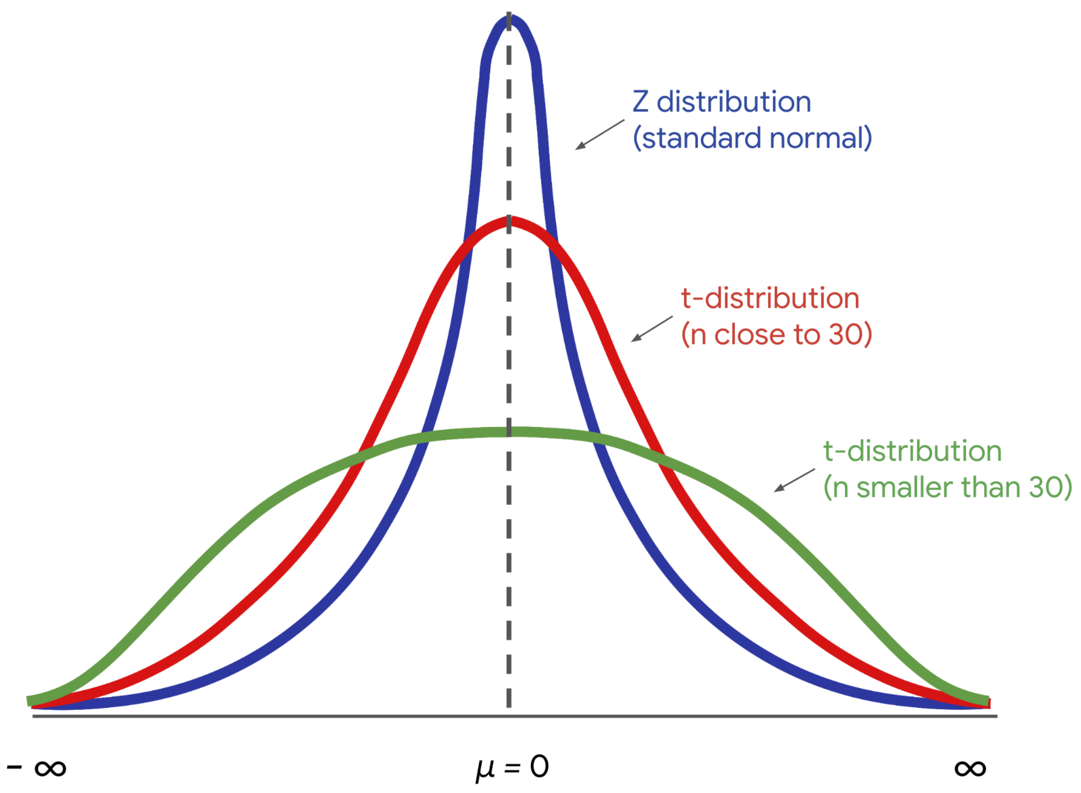
Large sample: Z-scores

For large sample sizes, you use **z-scores** to calculate the margin of error, just like you did earlier to estimate mean battery life for cell phones. This is because of the central limit theorem: for large sample sizes, the sample mean is approximately normally distributed. For a standard normal distribution, also called a **z-distribution**, you use z-scores to make calculations about your data.

Small sample: T-scores

For small sample sizes, you need to use a different distribution, called the **t-distribution**. Statistically speaking, this is because there is more uncertainty involved in estimating the standard error for small sample sizes. You don't need to worry about the technical details, which are beyond the scope of this course. For now, just know that if you're working with a small sample size, and your data is approximately normally distributed, you should use the t-distribution rather than the standard normal distribution. For a t-distribution, you use t-scores to make calculations about your data.

The graph of the t-distribution has a bell shape that is similar to the standard normal distribution. But, the t-distribution has bigger tails than the standard normal distribution does. The bigger tails indicate the higher frequency of outliers that come with a small dataset. As the sample size increases, the t-distribution approaches the normal distribution. When the sample size reaches 30, the distributions are practically the same, and you can use the normal distribution for your calculations.



Example: Mean emission levels

Now that you know a little bit about the t-distribution and t-scores, let's construct a confidence interval for a small sample size.

Context

Imagine you're a data professional working for an auto manufacturer. The company produces high performance cars that are sold around the world. Typically, the engines in these cars have high emission rates of carbon dioxide, or CO_2, which is a greenhouse gas that contributes to global warming. The engineering team has designed a new engine to reduce emissions for the company's best-selling car.

Goal

The goal is to keep emissions below 460 grams of CO_2 per mile. This will ensure the car meets emissions standards in every country it's sold in. Plus, the lower emissions rate is good for the environment, which will appeal to new customers.

Ask

The engineering team asks you to provide a reliable estimate of the emissions rate for the new engine. Due to production issues, there are only a limited number of engines available for testing. So, you’ll be working with a small sample size.

Sample

The engineering team tests a random sample of 15 engines and collects data on their emissions. The mean emission rate is 430 grams of CO_2 per mile, and the standard deviation is 35 grams of CO_2 per mile.

Your single sample may not provide the actual mean emissions rate for *every* engine. The population mean for emissions could be above or below 430 grams of CO_2 per mile. Even though you only have a small sample of engines, you can construct a confidence interval that likely includes the actual emission rate for a large population of engines. This will give your manager a better idea of the uncertainty in your estimate. It will also help the engineering team decide if they need to do more work on the engine to lower the emissions rate.

Construct the confidence interval

Let’s review the steps for constructing a confidence interval:

1. Identify a sample statistic.
2. Choose a confidence level.
3. Find the margin of error.
4. Calculate the interval.

Step 1: Identify a sample statistic

First, identify your sample statistic. Your sample represents the average emissions rate for 15 engines. You’re working with a sample *mean*.

Step 2: Choose a confidence level

Next, choose a confidence level. The engineering team requests that you choose a 95% confidence level.

Step 3: Find the margin of error

Your third step is to find the margin of error. For a small sample size, you calculate the margin of error by multiplying the t-score by the standard error.

The t-distribution is defined by a parameter called the degree of freedom. In our context, the degree of freedom is the sample size - 1, or 15-1 = 14. Given your degree of freedom and your confidence level, you can use a programming language like Python or other statistical software to calculate your t-score.

Based on a degree of freedom of 14, and a confidence level of 95%, your t-score is 2.145.

Now you can calculate the standard error, which measures the variability of your sample statistic.

Here’s the formula for the standard error of the mean that you’ve used before:

Standard Error (Means)

$\mathrm{SE}(\mathrm{x})=\mathrm{s}/\surd(\mathrm{n})$

$SE(x)=s/\surd(n)$

In the formula, the letter s refers to sample standard deviation, and the letter n refers to sample size.

Your sample standard deviation is 35, and your sample size is 15. The calculation gives you a standard error of about 9.04.

The margin of error is your t-score multiplied by your standard error. This is 2.145 * 9.04 = 19.39.

Step 4: Calculate the interval

Finally, calculate your confidence interval. The upper limit of your interval is the sample mean plus the margin of error. This is 430 + 19.39 = 449.39 grams of CO_2 per mile.

The lower limit is the sample mean minus the margin of error. This is 430 − 19.39 = 410.61 grams of CO_2 per mile.

You have a 95% confidence interval that stretches from 410.61 grams of CO_2 per mile to 449.39 grams of CO_2 per mile.

95 CI [410.61, 449.39]

The confidence interval gives the engineering team important information. The upper limit of your interval is below the target of 460 grams of CO_2 per mile. This result provides solid statistical evidence that the emissions rate for the new engine will meet emissions standards.

Note: Confidence intervals for small sample sizes only deal with population means, and not population proportions. The statistical reason for this distinction is rather technical, so you don’t need to worry about it for now.