Confidence Intervals

Overview of Course Progress:

- Gained knowledge in descriptive and inferential statistics.
- Learned basic rules of probability, including addition and multiplication rules.
- Explored probability distributions like binomial, Poisson, and normal distributions.
- Understood the sampling process, including different sampling methods.
- Learned about sampling distributions to estimate means and proportions.

Introduction to Confidence Intervals:

- A confidence interval is a range of values that describes the uncertainty surrounding an estimate.
- It is a frequentist concept, distinct from Bayesian credible intervals, which have different statistical definitions and procedures.
- Confidence intervals help describe the uncertainty of an estimate and are widely used in statistics and data science.

Importance of Confidence Intervals:

- Confidence intervals are a crucial tool in data-driven work for describing uncertainty in estimates.
- Common use cases include estimating the average return on investment for a stock portfolio, maintenance costs for machinery, customer registration percentages for rewards programs, and click-through rates on websites.
- Misinterpretation of confidence intervals can lead to false conclusions in studies, so understanding them correctly is essential.

Construction and Interpretation:

- The course will cover the procedure for constructing confidence intervals:
    - Identifying a sample statistic.
    - Choosing a confidence level.
    - Finding the margin of error.
    - Calculating the interval.
- Confidence intervals can be constructed for both means and proportions.

Practical Application:

- You'll learn how to use Python's SciPy stats module to construct a confidence interval for a point estimate of a population mean.

Relevance to Career:

- Confidence intervals are commonly used by data professionals and may be a part of your future job responsibilities.
- Understanding confidence intervals is important for job interviews and practical data analysis work.

Future Learning:

- While this course provides a foundation, ongoing learning about confidence intervals is encouraged as they are an active topic of discussion in the field of statistics and data science.

Confidence Intervals and Point Estimates

Point Estimates vs. Interval Estimates:

- **Point Estimate:** Uses a single value to estimate a population parameter. Example: Estimating the mean weight of a population of 10,000 penguins as 31 pounds based on a sample.
- **Interval Estimate:** Uses a range of values to estimate a population parameter. Example: Constructing a 95% confidence interval between 28-32 pounds for penguin weight.

2. Purpose of Confidence Intervals:

- Confidence intervals provide a way to express the uncertainty in an estimate, which is caused by the randomness inherent in sampling.
- Unlike point estimates, confidence intervals account for this uncertainty and give a more reliable estimate.

3. Components of a Confidence Interval:

- **Sample Statistic:** The point estimate derived from the sample. Example: A sample mean of 30 pounds for penguin weight.
- **Margin of Error:** Represents the maximum expected difference between the population parameter and the sample estimate. It is added and subtracted from the sample statistic to determine the confidence interval.
- **Confidence Level:** Describes the likelihood that the confidence interval will include the population parameter. Common confidence levels are 90%, 95%, and 99%.

4. Example Calculation:

- If the sample mean of penguin weight is 30 pounds and the margin of error is ±2 pounds, the confidence interval is from 28 to 32 pounds.
- The confidence level (e.g., 95%) indicates that if 100 random samples were taken, approximately 95 of those samples would produce a confidence interval containing the actual population mean.

5. Importance of Confidence Intervals:

- Confidence intervals are more informative than point estimates because they communicate the uncertainty in the estimate.
- This is crucial in decision-making, as it provides a range of possible values rather than a single, potentially misleading number.

6. Practical Application Example:

- For a fashion company estimating sales revenue, a point estimate might be "$1 million," while a confidence interval might be "$950,000 to $1,050,000 at a 95% confidence level."
- The confidence interval gives stakeholders more information to make informed decisions.

7. Communicating Confidence Intervals:

- As a data professional, it's important to ensure stakeholders understand how to interpret a confidence interval and the uncertainty it represents.

Correct and incorrect interpretations

Recently, you learned that data professionals use confidence intervals to help describe the uncertainty surrounding an estimate. To better understand your data, and effectively communicate your results to stakeholders, it’s important to know how to correctly interpret a confidence interval.

In this reading, we’ll review the correct way to interpret a confidence interval. We’ll also discuss some common forms of misinterpretation and how to avoid them.

Correct interpretation:

Example: mean weight

Let’s explore an example to get a better understanding of how to interpret a confidence interval. Imagine you want to estimate the mean weight of a population of 10,000 penguins. Instead of weighing every single penguin, you select a sample of 100 penguins. The mean weight of your sample is 30 pounds. Based on your sample data, you construct a 95% confidence interval between 28 pounds and 32 pounds.

95 CI [28, 32]

Interpret the confidence interval

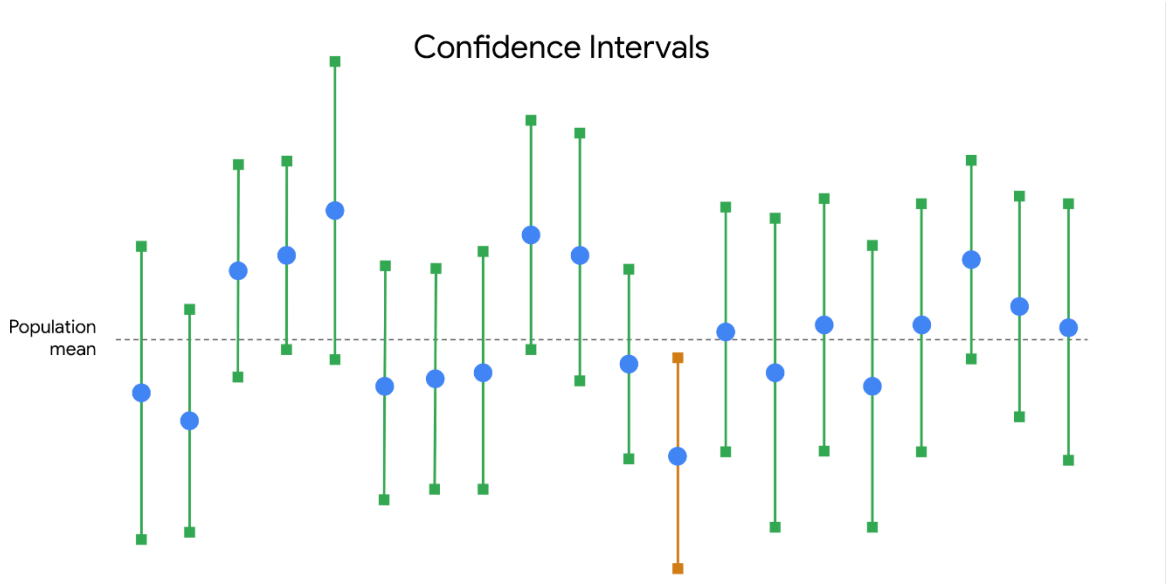
Earlier, you learned that the confidence level expresses the uncertainty of the estimation process. Let’s discuss what 95% confidence means from a more technical perspective.

Technically, 95% confidence means that if you take repeated random samples from a population, and construct a confidence interval for each sample using the same method, you can expect that 95% of these intervals will capture the population mean. You can also expect that 5% of the total will *not* capture the population mean.

The confidence level refers to the long-term success rate of the **method**, or the estimation process based on random sampling.

For the purpose of our example, let’s imagine that the mean weight of all 10,000 penguins is 31 pounds, although you wouldn’t know this unless you actually weighed every penguin. So, you take a sample of the population.

Imagine you take 20 random samples of 100 penguins each from the penguin population, and calculate a 95% confidence interval for each sample. You can expect that approximately 19 of the 20 intervals, or 95% of the total, will contain the actual population mean weight of 31 pounds. One such interval will be the range of values between 28 pounds and 32 pounds.



In practice, data professionals usually select one random sample and generate one confidence interval, which may or may not contain the actual population mean. This is because repeated random sampling is often difficult, expensive, and time-consuming. Confidence intervals give data professionals a way to quantify the uncertainty due to random sampling.

Incorrect interpretations
Now that you have a better understanding of how to properly interpret a confidence interval, let's review some common misinterpretations and how to avoid them.

Misinterpretation 1: 95% refers to the probability that the population mean falls within the constructed interval

One incorrect statement that is often made about a confidence interval at a 95% level of confidence is that there is a 95% probability that the population mean falls within the constructed interval.

In our example, this would mean that there's a 95% chance that the mean weight of the penguin population falls in the interval between 28 pounds and 32 pounds.

This is incorrect. The population mean is a constant.

Like any population parameter, the population mean is a constant, not a random variable. While the value of the sample mean varies from sample to sample, the value of the population mean does not change. The probability that a constant falls within any given range of values is always 0% or 100%. It either falls within the range of values, or it doesn't.

For example, any given random sample of 100 penguins may have a different mean weight: 32.8 pounds, 27.3 pounds, 29.6 pounds, and so on. You can use a sampling distribution to assign a specific probability to each of your sample means because these are random variables. However, the population mean weight is considered a constant. In our example, if you weigh all 10,000 penguins, you'll find that the population mean is 31 pounds. This value is fixed, and does not vary from sample to sample.

Sample Mean (100 penguins)	**Population Mean (10,000 penguins)**
32.8 lbs	31 lbs
27.3 lbs	31 lbs

Sample Mean (100 penguins)	Population Mean (10,000 penguins)
29.6 lbs	31 lbs

So, it's not strictly correct to say there is a 95% chance that your confidence interval captures the population mean because this implies that the population mean is variable. Intervals change from sample to sample, but the value of the population mean you're trying to capture does not.

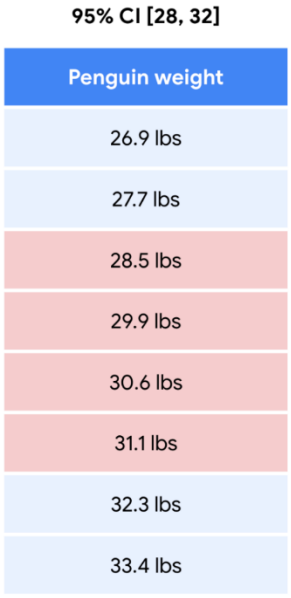
What you can say is that if you take repeated random samples from the population, and construct a confidence interval for each sample using the same method, you can expect 95% of your intervals to capture the population mean.

Pro tip: Remember that a 95% confidence level refers to the success rate of the estimation process.

Misinterpretation 2: 95% refers to the percentage of data values that fall within the interval

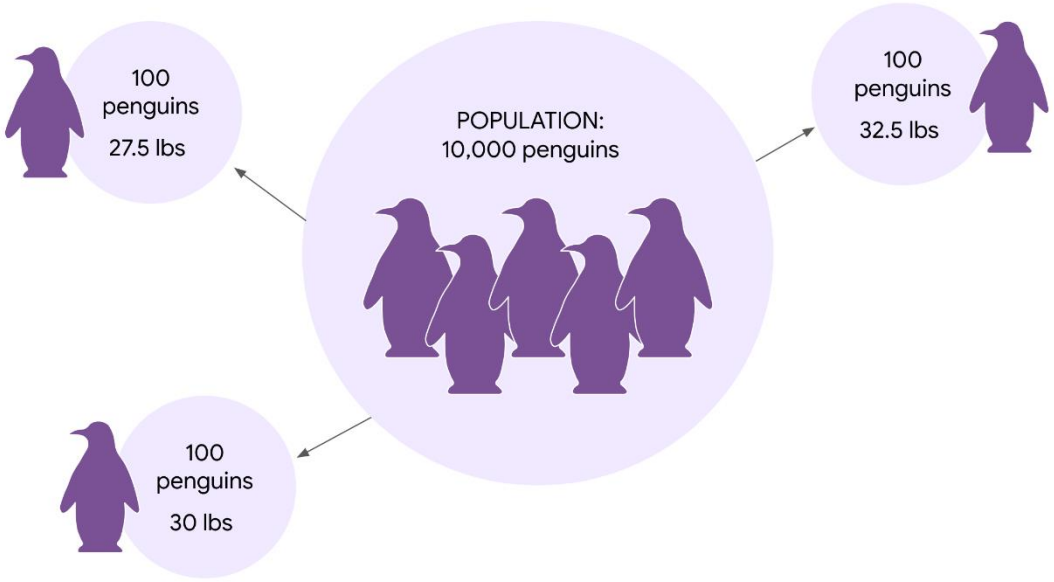
Another common mistake is to interpret a 95% confidence interval as saying that 95% of all of the data values in the population fall within the interval. This is not necessarily true. A 95% confidence interval shows a range of values that likely includes the actual population mean. This is *not* the same as a range that contains 95% of the data values in the population.

For example, your 95% confidence interval for the mean penguin weight is between 28 pounds and 32 pounds. It may not be accurate to say that 95% of all weight values fall within this interval. It's possible that over 5% of the penguin weights in the population are outside this interval—either less than 28 pounds or greater than 32 pounds.



Misinterpretation 3: 95% refers to the percentage of sample means that fall within the interval

A third common misinterpretation is that a 95% confidence interval implies that 95% of all possible sample *means* fall within the range of the interval. This is not necessarily true. For example, your 95% confidence interval for mean penguin weight is between 28 pounds and 32 pounds. Imagine you take repeated samples of 100 penguins and calculate the mean weight for each sample. It's possible that *over* 5% of your sample means will be less than 28 pounds or greater than 32 pounds.



Key takeaways
Knowing how to correctly interpret confidence intervals will give you a better understanding of your estimate, and help you share useful and accurate information with stakeholders. You may need to explain the common misinterpretations too, and why they're incorrect. You don't want your stakeholders to base their decisions on a misinterpretation. Understanding how to effectively communicate your results to stakeholders is a key part of your job as a data professional.

Construct Confidence Intervals

Constructing Confidence Intervals for Proportions

1. **Example Scenario:**

- **Context:** Polling for an upcoming election between Tiffany Davis and Maya Cruz.
- **Sample:** 100 voters out of 100,000 total voters.
- **Result:** 55% prefer Davis, 45% prefer Cruz.

2. Point Estimate vs. Confidence Interval:

- **Point Estimate:** 55% of voters prefer Davis.
- **Confidence Interval:** Provides a range to express the uncertainty around the point estimate.

3. Steps to Construct a Confidence Interval for a Proportion:

1. **Identify the Sample Statistic:**
    - **Sample Proportion (p):** 0.55 (55%).
2. **Choose a Confidence Level:**
    - **Confidence Level:** 95% (commonly used in polls).
3. **Find the Margin of Error:**
    - **Z-Score for 95% Confidence Level:** 1.96.
    - **Standard Error (SE):** Measures variability. Calculated as

$$\text{SE} = \sqrt{\frac{p \times (1-p)}{n}}$$

Where $p = 0.55$ and $n = 100$:

$$\text{SE} = \sqrt{\frac{0.55 \times (1-0.55)}{100}} \approx 0.05$$

- **Margin of Error (ME):**

$$\text{ME} = \text{Z-Score} \times \text{SE} = 1.96 \times 0.05 = 0.098$$

Calculate the Confidence Interval:

- **Upper Limit:**

$$\text{Upper Limit} = p + \text{ME} = 0.55 + 0.098 = 0.648 \text{ or } 64.8\%$$

- **Lower Limit:**

$$\text{Lower Limit} = p - \text{ME} = 0.55 - 0.098 = 0.452 \text{ or } 45.2\%$$

- **Confidence Interval:** 45.2% to 64.8%.

4. Interpretation:

- The confidence interval provides a range within which the true proportion is likely to fall with the chosen confidence level (95%).
- In this case, the interval includes values below 50%, indicating that Davis could potentially lose the election.

5. Impact of Sample Size:

- **Larger Sample Size:** Results in a narrower confidence interval and a more accurate estimate.
- **Example:** With 1,000 voters, if the sample proportion is 54%, the 95% confidence interval might be from 50.9% to 57.1%.
- **Narrower Interval:** More precise estimate, e.g., 6.2 percentage points for 1,000 voters vs. 19.6 percentage points for 100 voters.

6. Conclusion:

- **Confidence Intervals:** Help communicate the uncertainty and provide a range of plausible values for the population proportion.
- **Application:** Important for making informed decisions and understanding the reliability of the sample estimate.

Confidence Interval for a Mean

Example Scenario:

- **Context:** Analyzing battery life for a new cell phone.
- **Sample Data:** 100 phones tested.
- **Sample Mean (M):** 20.5 hours.
- **Sample Standard Deviation (s):** 1.7 hours.
- **Population Standard Deviation (σ):** 1.5 hours.

Steps to Construct a Confidence Interval for a Mean:

1. Identify the Sample Statistic:

- **Sample Mean (M):** 20.5 hours.

2. Choose a Confidence Level:

- **Confidence Level:** 95% (default for many studies).
- **Z-Score for 95% Confidence Level:** 1.96.

3. Find the Margin of Error:

- **Standard Error (SE):**

$$\text{SE} = \frac{\sigma}{\sqrt{n}}$$

Where $\sigma = 1.5$ (population standard deviation) and $n = 100$ (sample size):

$$\text{SE} = \frac{1.5}{\sqrt{100}} = 0.15$$

- Margin of Error (ME):

$$\text{ME} = \text{Z-Score} \times \text{SE} = 1.96 \times 0.15 = 0.294$$

4. Calculate the Confidence Interval:

- Upper Limit:

$$\text{Upper Limit} = M + \text{ME} = 20.5 + 0.294 = 20.794 \text{ hours (about 20 hours and 48 m}$$

- Lower Limit:

$$\text{Lower Limit} = M - \text{ME} = 20.5 - 0.294 = 20.206 \text{ hours (about 20 hours and 12 m}$$

- Confidence Interval: 20 hours and 12 minutes to 20 hours and 48 minutes.

5. Interpretation:

- **95% Confidence Interval:** The interval suggests that with 95% confidence, the true population mean battery life is between 20 hours and 12 minutes and 20 hours and 48 minutes.

- **Marketing Claim:** Since the entire interval is above 20 hours, the marketing team can confidently advertise the phone's battery life as at least 20 hours.

Additional Analysis with 99% Confidence Level:

- **Z-Score for 99% Confidence Level:** 2.58.
- **Margin of Error (ME) with 99% Confidence Level:**

  $$ME=2.58\times0.15=0.387$$

- **Upper Limit:**

  $$\text{Upper Limit}=20.5+0.387=20.887 \text{ hours (about 20 hours and 53 minutes)}$$

- **Lower Limit:**

  $$\text{Lower Limit}=20.5-0.387=20.113 \text{ hours (about 20 hours and 7 minutes)}$$

- **99% Confidence Interval:** 20 hours and 7 minutes to 20 hours and 53 minutes.

6. Conclusion:

- **Wider Interval:** As the confidence level increases from 95% to 99%, the interval widens to include more of the possible values, reflecting greater certainty but less precision.
- **Implication:** The wider interval with a 99% confidence level still shows that the phone's battery life is above 20 hours, providing even more assurance for the marketing claim.

Note: When the population standard deviation is unknown, the sample standard deviation is used in place of the population standard deviation, and the t-distribution is used instead of the normal distribution.

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.

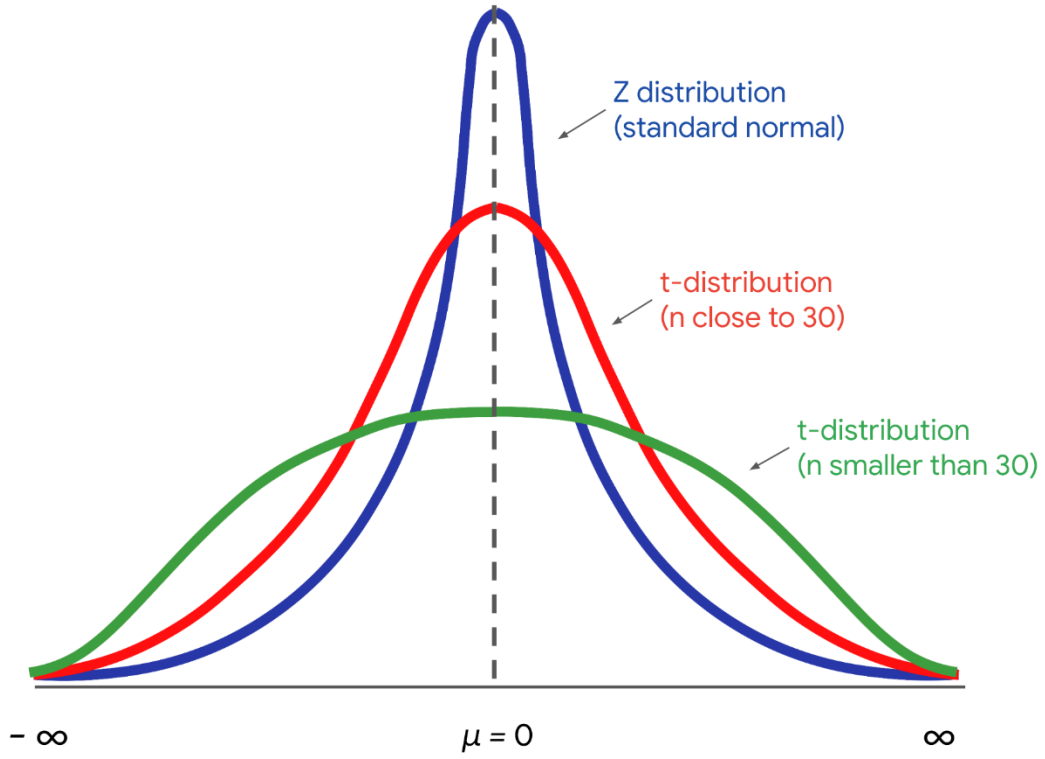
Large sample: Z-scores

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)

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

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.

Key takeaways
As a data professional, you'll work with both large and small sample sizes. Although large samples give more precise estimates than small samples, collecting a small sample is usually less expensive and time-consuming than collecting a large sample. Knowing how to construct confidence intervals for different sample sizes will help you manage any dataset that you may encounter in your future career.

Glossary

Confidence interval: A range of values that describes the uncertainty surrounding an estimate

Confidence level: A measure that expresses the uncertainty of the estimation process **Interval**: A sample statistic plus or minus the margin of error

Interval estimate: A calculation that uses a range of values to estimate a population parameter

Lower limit: When constructing an interval, the calculation of the sample means minus the margin of error

Margin of error: The maximum expected difference between a population parameter and a sample estimate

Method: A function that defines and performs behaviors like computation

Point estimate: A calculation that uses a single value to estimate a population parameter

Standard error of the mean: The sample standard deviation divided by the square root of the sample size

Standard error of the proportion: The square root of the sample proportion times one minus the sample proportion divided by the sample size

Upper limit: When constructing an interval, the calculation of the sample means plus the margin of error