Infer population parameters with the central limit theorem

Recently, you learned about the central limit theorem and how it can help you work with a wide variety of datasets. Data professionals use the central limit theorem to estimate population parameters for data in economics, science, business, and many other fields.

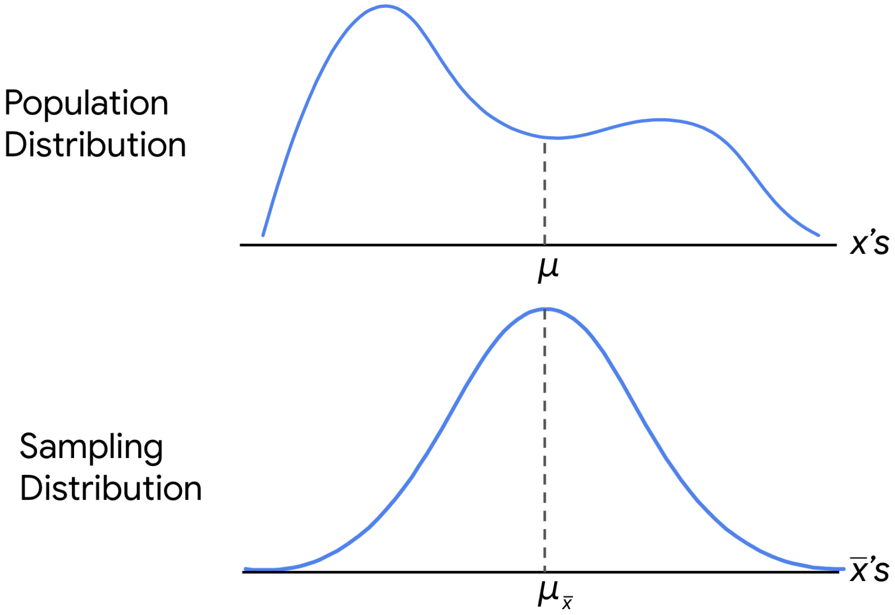
In this reading, you'll learn more about the central limit theorem and how it can help you estimate the population mean for different types of data. We'll go over the definition of the theorem, the conditions that must be met to apply the theorem, and check out an example of the theorem in action.

The central limit theorem

Definition

The **central limit theorem** states that the sampling distribution of the mean approaches a normal distribution as the sample size increases. In other words, as your sample size increases, your sampling distribution assumes the shape of a bell curve. And, as you sample more observations from a population, the sample mean gets closer to the population mean. If you take a large enough sample of the population, the sample mean will be roughly equal to the population mean.

For example, imagine you want to estimate the average weight of a certain class of vehicle, like light-duty pickup trucks. Instead of weighing millions of pickup trucks, you can get data on a representative sample of pickup trucks. If your sample size is large enough, the mean weight of your sample will be roughly equal to the mean weight of the population (adhering to the law of large numbers).



Note: The central limit theorem holds true for any population. You don't need to know the shape of your population distribution in advance to apply the theorem—the distribution could be bell-shaped, skewed, or have another shape. If you collect enough samples of sufficient size, the shape of the distribution of their means will follow a normal distribution.

Conditions

In order to apply the central limit theorem, the following conditions must be met:

- **Randomization***:* Your sample data must be the result of random selection. Random selection means that every member in the population has an equal chance of being chosen for the sample.
- **Independence***:* Your sample values must be independent of each other. Independence means that the value of one observation does not affect the value of another observation. Typically, if you know that the individuals or items in your dataset were selected randomly, you can also assume independence.
    - **10%***:* To help ensure that the condition of independence is met, your sample size should be no larger than 10% of the total population *when the sample is drawn without replacement* (which is usually the case).
        - **Note**: In general, you can sample with or without replacement. When a population element can be selected only one time, you are sampling without replacement. When a population element can be selected more than one time, you are sampling with replacement. You'll learn more about this topic later on in the course.
- **Sample size**: The sample size needs to be sufficiently large.

Let's discuss the sample size condition in more detail. There is no exact rule for how large a sample size needs to be in order for the central limit theorem to apply. The answer depends on the following factors:

- **Requirements for precision**. The larger the sample size, the more closely your sampling distribution will resemble a normal distribution, and the more precise your estimate of the population mean will be.
- **The shape of the population***.* If your population distribution is roughly bell-shaped and already resembles a normal distribution, the sampling distribution of the sample mean will be close to a normal distribution even with a small sample size.

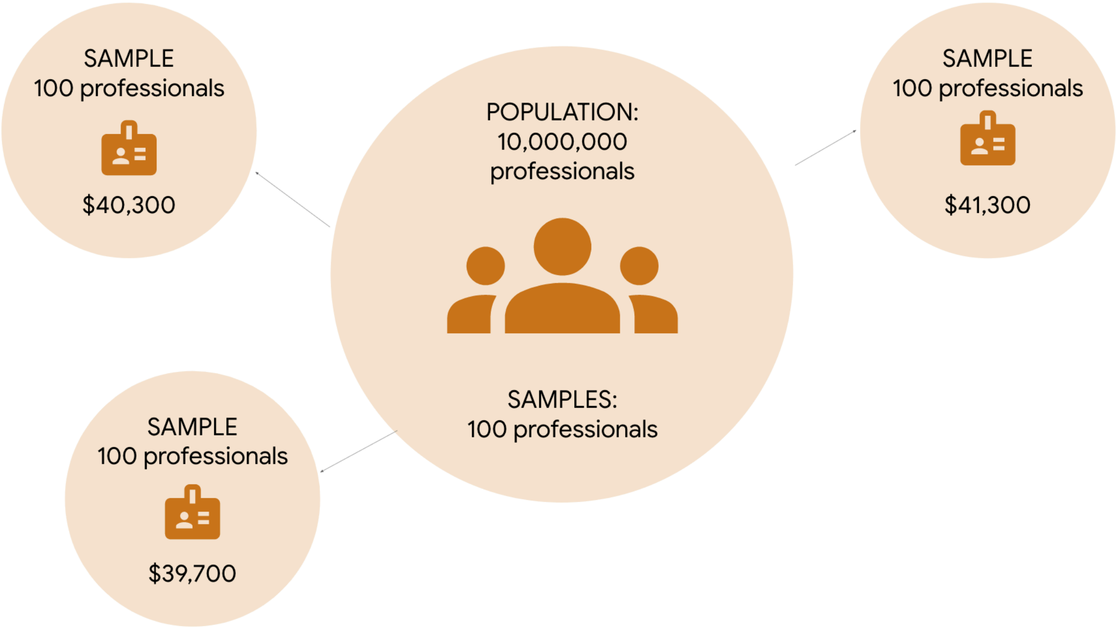
In general, many statisticians and data professionals consider a sample size of 30 to be sufficient when the population distribution is roughly bell-shaped, or approximately normal. However, if the original population is not normal—for example, if it's extremely skewed or has lots of outliers—data professionals often prefer the sample size to be a bit larger. Exploratory data analysis can help you determine how large of a sample is necessary for a given dataset.

Example: Annual salary

Let's explore an example to get a better idea of how the central limit theorem works.

Imagine you're studying annual salary data for working professionals in a large city like Buenos Aires, Cairo, Delhi, or Seoul. Let's say the professional population you're interested in includes 10 million people. You want to know the average annual salary for a professional living in the city. However, you don't have the time or money to survey millions of professionals to get complete data on every salary.

Instead of surveying the entire population, you collect survey data from repeated random samples of 100 professionals. Using this data, you calculate the mean annual salary in dollars for your first sample: $40,300. For your second sample, the mean salary is: $41,100. You survey a third sample. The mean salary is $39,700. And so on. Due to sampling variability, the mean of each sample will be slightly different.



In theory, you could take a very large sample and increase the sample size until you've surveyed all 10 million people about their salary. The central limit theorem says that as your sample size increases, the shape of your sampling distribution will increasingly resemble a bell curve.

If you take a large enough sample from the population, the mean of your sampling distribution will be roughly equal to the population mean. From this sample of the population, you can precisely estimate the average annual salary for the entire professional population.

Note: In practice, data professionals usually take a single sample. The specific sample size they choose depends on factors like budget, time, resources, and the desired level of confidence for their estimate.