



Reading: The sampling distribution of the mean

Recently, you've learned about how data professionals use sample statistics to estimate population parameters. For example, a data professional might estimate the mean time customers spend on a retail website, or the mean salary of all the people who work in the entertainment industry.

In this reading, you'll learn more about the concept of sampling distribution and how it can help you represent the possible outcomes of a random sample. We'll also discuss how the sampling distribution of the sample mean can help you estimate the population mean.

Sampling distribution of the sample mean

A **sampling distribution** is a probability distribution of a sample statistic. Recall that a **probability distribution** represents the possible outcomes of a random variable, such as a coin toss or a die roll. In the same way, a sampling distribution represents the possible outcomes for a sample statistic. Sample statistics are based on randomly sampled data, and their outcomes cannot be predicted with certainty. You can use a sampling distribution to represent statistics such as the mean, median, standard deviation, range, and more.

Typically, data professionals compute sample statistics like the mean to estimate the corresponding population parameters.

Suppose you want to estimate the mean of a population, like the mean height of a group of humans, animals, or plants. A good way to think about the concept of sampling distribution is to imagine you take repeated samples from the population, each with the same sample size, and compute the mean for each of these samples. Due to sampling variability, the sample mean will vary from sample to sample in a way that cannot be predicted with certainty. The distribution of all your sample means is essentially the sampling distribution. You can display the distribution of sample means on a histogram. Statisticians call this the sampling distribution of the mean.

Note: In practice, due to limited time and resources, data professionals typically collect a single sample and calculate the mean of that sample to estimate the population mean.

Let's explore an example to get a more concrete idea of the sampling distribution of the mean.

Example: Mean length of lake trout

You are a data professional working with a team of environmental scientists. Your team studies the effects of water pollution on fish species. Currently, your team is researching the effects of pollution on the trout population in Lake Superior, one of the Great Lakes in North America. As part of this

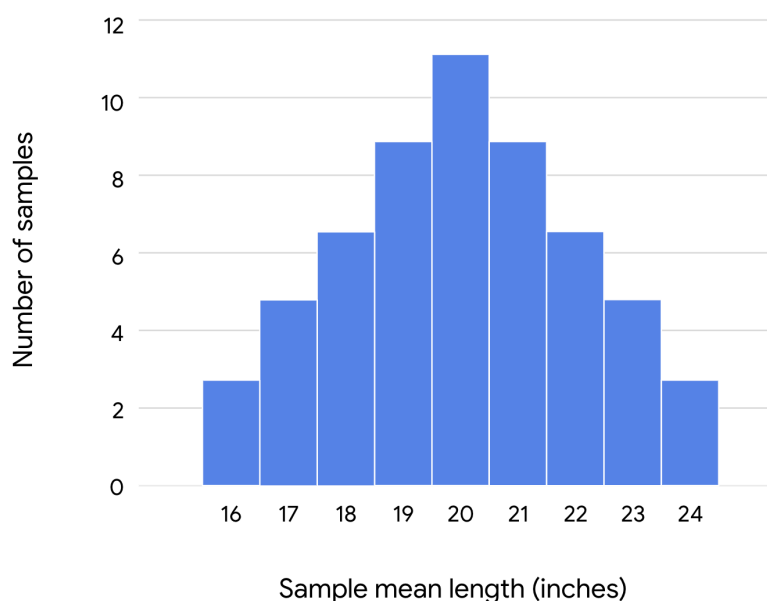
research, they ask you to estimate the mean length of a trout. Let's say there are 10 million trout in the lake. Instead of collecting and measuring millions of trout, you take sample data from the population.

Let's say you take repeated simple random samples of 100 trout each from the population. In other words, you randomly choose 100 trout from the lake, measure them, and then repeat this process with a different set of 100 trout. For your first sample of 100 trout, you find that the mean length is 20.2 inches. For your second sample, the mean length is 20.5 inches. For your third sample, the mean length is 19.7 inches. And so on. Due to sampling variability, the mean length will vary randomly from sample to sample.

For the purpose of this example, let's assume that the true mean length of a trout in this population is 20 inches. Although, in practice, you wouldn't know this unless you measured every single trout in the lake.

Each time you take a sample of 100 trout, it's likely that the mean length of the trout in your sample will be close to the population mean of 20 inches, but not exactly 20 inches. Every once in a while, you may get a sample full of shorter than average trout, with a mean length of 16 inches or less. Or, you might get a sample full of longer than average trout, with a mean length of 24 inches or more.

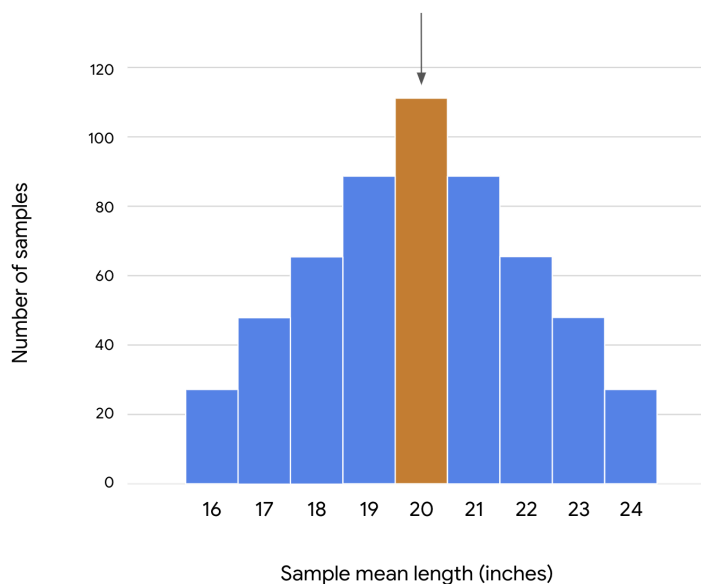
You can use a sampling distribution to represent the frequency of all your different sample means. For example, if you take 10 simple random samples of 10 trout each from the population, you can show the sampling distribution of the mean as a histogram. The most frequently occurring value in your sample data will be around 20 inches. The values that occur least frequently will be the more extreme lengths, such as 16 inches or 24 inches.



As you increase the size of a sample, the mean length of your sample data will get closer to the mean length of the population. If you sampled the entire population—in other words, if you actually

measured every single trout in the lake—your sample mean would be the same as the population mean.

However, you don't need to measure millions of fish to get an accurate estimate of the population mean. If you take a large enough sample size from the population—say, 1000 trout—your sample mean will be a precise estimate of the population mean (20 inches).



Standard error

You can also use your sample data to estimate how precisely the mean length of any given sample represents the population mean.

This is useful to know because the sample mean varies from sample to sample, and any given sample mean is likely to differ from the true population mean. For example, the mean length of the trout population might be 20 inches. The mean length for any given sample of trout might be 20.2 inches, 20.5 inches, 19.7 inches, and so on.

Data professionals use the standard deviation of the sample means to measure this variability. In statistics, the standard deviation of a sample statistic is called the **standard error**. The standard error provides a numerical measure of sampling variability. The standard error of the mean measures variability among all your sample means. A larger standard error indicates that the sample means are more spread out, or that there's more variability. A smaller standard error indicates that the sample means are closer together, or that there's less variability.

In practice, using a single sample of observations, you can apply the following formula to calculate the estimated standard error of the sample mean: s / \sqrt{n} . In the formula, s refers to the sample standard deviation, and n refers to the sample size.

For example, in your study of trout lengths, imagine that a sample of 100 trout has a mean length of 20 inches and a standard deviation of 2 inches. You can calculate the estimated standard error by dividing the sample standard deviation, 2, by the square root of the sample size, 100:

$$2 \div \sqrt{100} = 2 \div 10 = 0.2$$

This means you should expect that the mean length from one sample to the next will vary with a standard deviation of about 0.2 inches.

The standard error helps you understand the precision of your estimate. In general, you can have more confidence in your estimates as the sample size gets larger and the standard error gets smaller. This is because, as your sample size gets larger, the sample mean gets closer to the population mean.

Key takeaways

Estimating population parameters through sampling is a powerful form of statistical inference. Sampling distributions describe the uncertainty associated with a sample statistic, and help you make proper statistical inferences. This is important because stakeholder decisions are often based on the estimates you provide.
