

Concept of a Sampling Distribution

Statistical inference involves using sample data to draw conclusions about a population. For example, if we want to say something about the average income of all St. John's residents (our **population**), it would be difficult and time-consuming to ask every resident what their income is. Instead we could take a random sample of 100 (for example) residents, and use the average income of this group to **estimate** the population average income. If we let the population mean (or average) be denoted by μ (mu), and the sample mean be denoted by \bar{x} , then \bar{x} would be used as our estimate of μ . But to use \bar{x} to make some inferential statement about μ , we need some idea of how \bar{x} behaves, in general, as an estimate. One way to do this is to consider the **sampling distribution** of the estimate. As we will briefly illustrate here, the concept of a sampling distribution is a long-run, or repeated sampling one.

Although we will describe the sampling distribution of \bar{x} , the procedure applies to any **statistic**, or quantity calculated from sample data: the sample median, sample standard deviation, etc.

Suppose we have a huge pot of numbers that represents that population distribution of interest. This may be a normal distribution, or something with a very non-normal distribution.

1. Draw a random sample from the pot, and find the sample mean. Call it \bar{x}_1 .
2. Toss the values back in the pot. Pull out another sample, find its mean. Call it \bar{x}_2 .
3. Repeat these steps 1000's of times, get 1000's of \bar{x} 's.
4. Draw a histogram (barchart) of these \bar{x} values.

The histogram you get in step 4 is the **sampling distribution** of \bar{x} . It is this distribution that is used to derive confidence intervals and test statistics used in hypothesis testing.

A very important result in statistics comes from the idea of sampling distributions. The result is the **Central Limit Theorem**:

Central Limit Theorem: If x is a random variable that comes from a distribution with mean μ and standard deviation σ , then for a "large enough" sample size n , \bar{x} is approximately normally distributed with mean μ and standard deviation σ/\sqrt{n} .

This tells us that, even if we're dealing with a complicated population distribution, the sample mean still behaves as if comes from a normal distribution.

So, how large does the sample size have to be? The rule of thumb is that we need $n \geq 30$ in order for the Central Limit Theorem to hold.