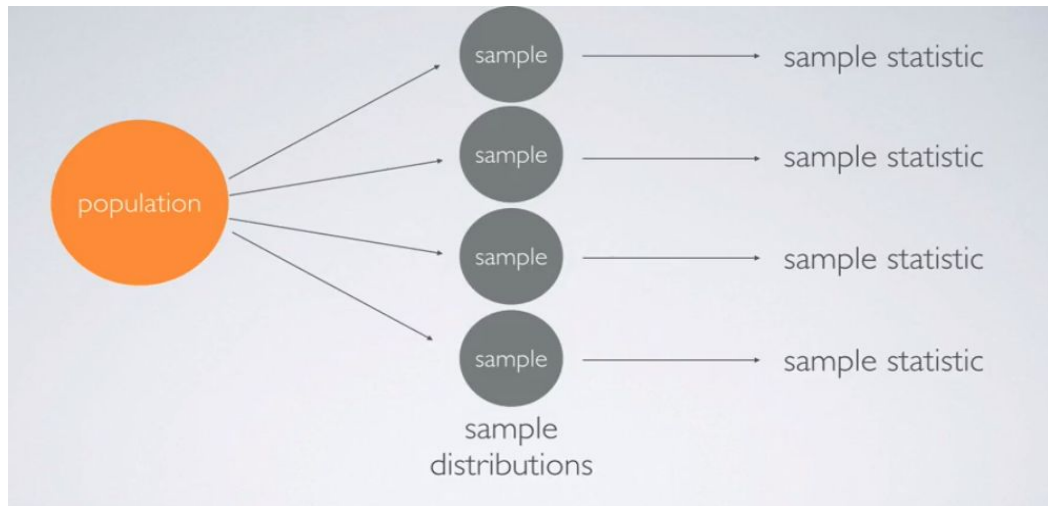


## Notes on Sampling Vs Bootstrap Distributions

Made by Aparna Shastry

**Sampling Distribution:** Consider a population of size  $N$ , and samples of size  $n$  taken from that population. There are two ways to take samples (also known as sampling the population): with replacement or without replacement. Without replacement sampling requires that  $n < 0.1N$  or preferably,  $< 0.05N$



Sample statistics in this case is mean. CLT says how the mean (also called sample statistic) of these samples is distributed. Note that sample distribution and sampling distributions are different. Sample distribution is same as original population distribution, sampling distribution is distribution of mean of samples taken from the population.

**Central Limit Theorem (CLT):** The distribution of sample statistics is nearly normal, centered at the population mean, and with a standard deviation equal to the population standard deviation divided by square root of the sample size.

$$\bar{x} \sim N \left( \text{mean} = \mu, SE = \frac{\sigma}{\sqrt{n}} \right)$$

↓      ↓      ↓  
Shape   Center   Spread

### Conditions for the CLT:

1. **Independence:** Sampled observations must be independent.
  - ▶ random sample/assignment
  - ▶ if sampling without replacement,  $n < 10\%$  of population
2. **Sample size/skew:** Either the population distribution is normal, or if the population distribution is skewed, the sample size is large (rule of thumb:  $n > 30$ ).

If population is not nearly normal, the more skewed it is, the larger  $n$  we need, for the CLT to apply.

**Bootstrap Distribution:** We don't have access to population almost always. We have access to one sample of the population. Then we do bootstrapping to simulate the effect of taking multiple samples from the population.

### bootstrapping scheme

- (1) take a bootstrap sample - a random sample taken **with replacement** from the original sample, of the same size as the original sample
- (2) calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples

Distribution of sample statistic of bootstrap sample taken from a sample with replacement, of same size as original sample

### bootstrapping limitations

- ▶ Not as rigid conditions as CLT based methods
- ▶ If the bootstrap distribution is extremely skewed or sparse, the bootstrap interval might be unreliable
- ▶ A representative sample is still required — if the sample is biased, the estimates resulting from this sample will also be biased.

### Summary:

- Sampling Distribution is created by sampling with or without from the population. The statistics under consideration is mean or standard deviation
- Bootstrap Distribution is created by sampling with replacement from one sample of the population. The statistics under consideration can be mean, median, proportion or standard deviation
- Both are distributions of sample statistics
- Both are useful in estimating true population statistics.

Images: Screenshots from slides of [Duke university coursera course on inferential statistics](#)

(Copyright: Please do not distribute without author's permission)