ST 516: Foundations of Data Analytics

The Sample Mean

The Sample Mean

The Central Limit Theorem

Sampling Distributions

Recall that the sampling distribution of a statistic (or random variable) is the distribution of that statistic based on all possible samples of a fixed size from the population of interest.

- We've been using the computer to generate these repeated samples, so that you can actually look at some sampling distributions
- Of course in practice we only observe one sample, but computer simulations can be very useful for understanding how we expect statistics based on that one sample to behave
- And the sampling distribution of a statistic allows us to quantify probabilities associate with that statistic

The Sample Mean

The sample mean (the arithmetic average of a sample of data) is an important statistic, since it's often the first thing that people think to calculate to summarize a sample.

The sample mean has some appealing properties that make it a very useful summary:

1. The expected value of the sample mean is equal to the mean of the population from which the sample is obtained.

2. The variance of the sample mean is equal to the variance of the population divided by the sample size.

3. And, there is a very important result in Statistics that tells us the shape of the sampling distribution of the sample mean when the sample size is large.

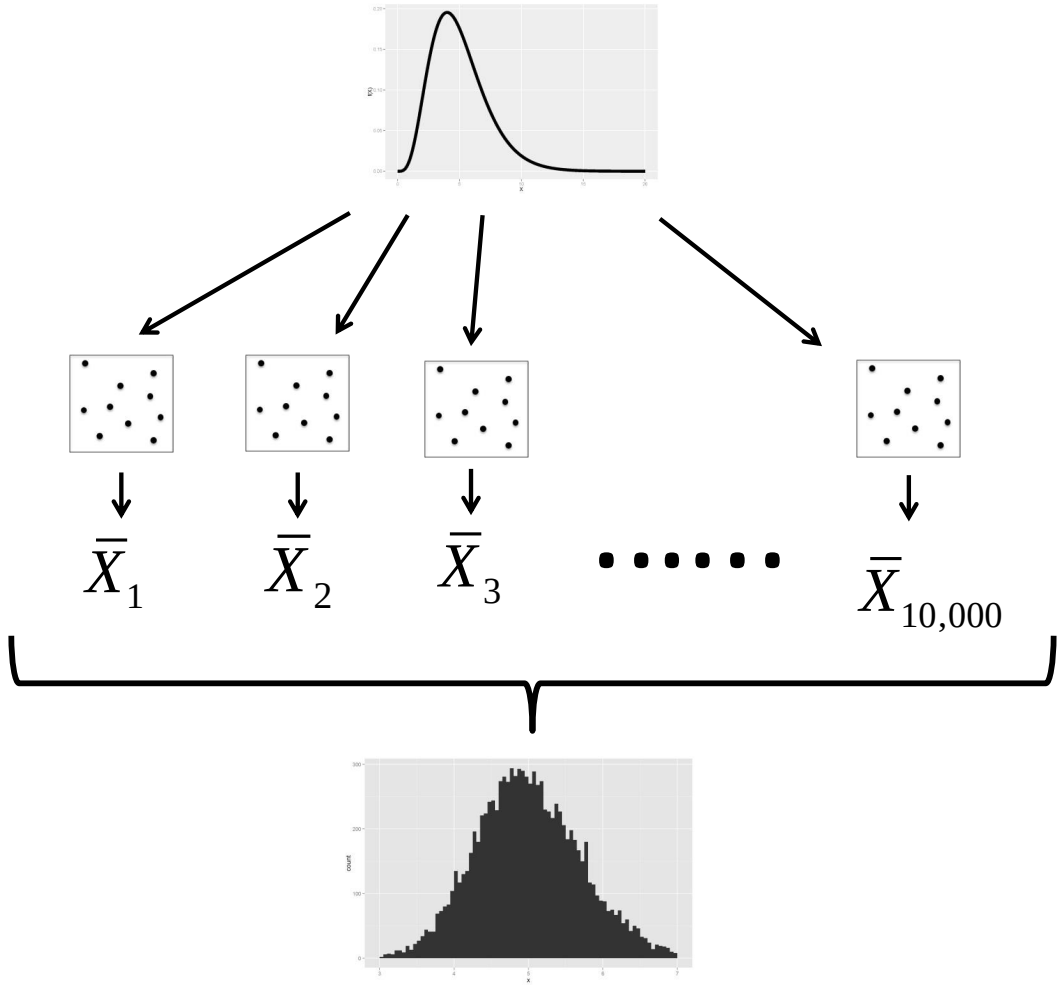
The Central Limit Theorem

The **Central Limit Theorem** tells us about the sampling distribution of the sample mean.

Let $X_1, X_2, \ldots, X_n$ denote an independent sample of random variables from a population that has unknown mean μ and unknown variance σ^2.Then, for large sample sizes, the sampling distribution of the sample mean, $\overline{X}$, is approximately Normal with mean μ and variance σ^2/n.

The beauty of this theorem is that it holds true *regardless of the shape* of the population distribution from which the sample is taken.

The Central Limit Theorem

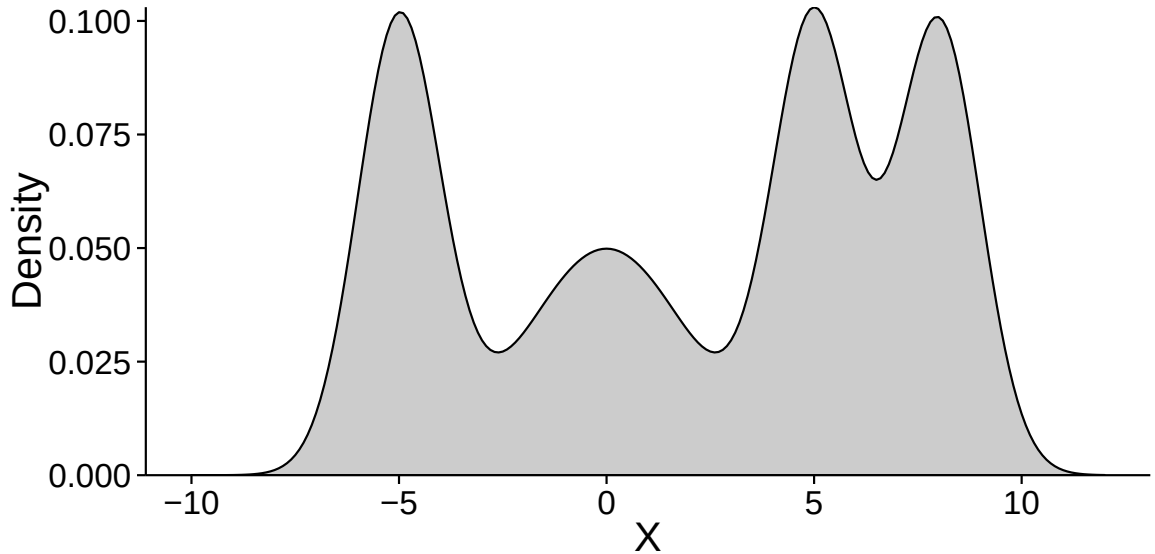


The Central Limit Theorem

A key question regarding the Central Limit Theorem (CLT) is: *How large does the sample size have to be for the CLT to hold?*.

- Technically, the CLT has the word "limit" in it, which means that the sample size has to be infinite for approximation in the theorem to be exact.
- Most of the time, however, the approximation works really well for sample sizes larger than about 30 (when the population distribution is particularly skewed—or asymmetric—the sample size may need to be larger).
- The CLT holds for population distributions that are pretty dissimilar to the Normal distribution.

A Non-Normal Distribution



The Central Limit Theorem