ST 516: Foundations of Data Analytics

Inference

Goals of Data Analysis

Statistical Inference

- From Sampling to Inference

Goals of Data Analysis

Recall that the goals of data analysis fall into three categories:

1. Exploration and description
2. Inference (hypothesis testing and estimation)
3. Prediction

In the remainder of this course, we'll be talking mostly about inference, although we'll emphasize exploration and description also.

Statistical Inference

Statistical inference is the process of learning—or inferring—properties of a population distribution from properties of a sample from that distribution.

For example, to learn about a population mean, μ, we'll use a sample mean, $\overline{X}$, and to learn about a population median, M, we'll use a sample median, m.

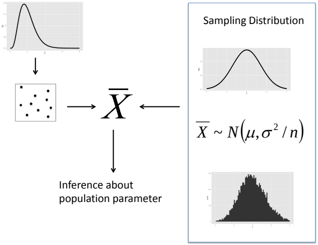
Because of the variability inherent in all data, and because we only typically look at a small portion of the data from a population (i.e., a sample), what we learn about population values will be *uncertain*. A primary goal of statistical inference is to quantify that uncertainty.

Statistical Inference

The language that we use to describe uncertainty related to statistical inference is probability.

- Remember that the sampling distribution captures the variation of the estimate of interest, because it is based on the notion of calculating the estimate from all possible samples of a fixed size.
- And remember, the sampling distribution is just a probability distribution based on the long-run relative frequency of an estimate.
- What we need to discuss now is how we can use an estimate from *a single sample* to learn something about the population distribution from which the sample was obtained.

From Sampling to Inference



$\bar{X}$

Inference about population parameter

Sampling Distribution

$\bar{X} \sim N(\mu, \sigma^2 / n)$

In the case of the sample mean, it's straightforward to relate properties of the sampling distribution of the sample mean to properties of the population distribution from which the sample was taken:

- The mean of the sampling distribution, also called the expected value of the statistic, is equal to the mean of the population.
- The variance of the sampling distribution is equal to the population variance divided by the sample size.

From Sampling to Inference

In many other cases, there are results from statistical theory that allow us to use properties of the sampling distribution of a particular estimate to learn about properties of the population distribution from which the sample was taken.

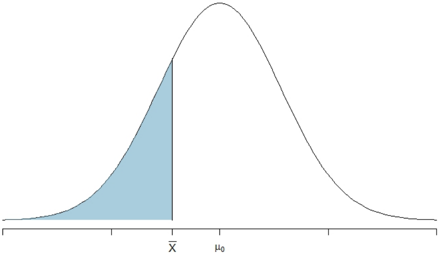
- For example, you'll see how a statistic related to the quantity $\overline{X}_1 - \overline{X}_2$ can be used to learn about $\mu_1 - \mu_2$, where $(\overline{X}_1, \overline{X}_2)$ and (μ_1, μ_2) are sample means and population means, respectively, from two populations.

A key idea underlying these connections is that the uncertainty in our estimates will be quantified by the variation in the sampling distributions in those estimates. So, we'll be able to make probability statements about our estimates by using areas under the curve of the corresponding sampling distribution.

Using a Sampling Distribution

The Central Limit Theorem tells us that for large samples, the sampling distribution of the sample mean is approximately Normally shaped, with mean and variance μ and σ^2/n, where μ and σ^2 are the population mean and variance, respectively.

For example, when we hypothesize that a population mean is equal to some specified value, say μ_0, we can ask about how unusual $\overline{X}$ is relative to that hypothesis, by looking at the sampling distribution.


$\overline{X}$
μ_0