Overview

Estimation and hypothesis testing build upon probability. Specifically, they are built on probability distributions, which in turn depend on the concept of random variables and this unit assumes you are familiar with these topics.

- Khan Academy resource on random variables and distributions

Scientists use statistical models to make general statements about real-world observations. Many types of data collected from many different sources often follow recognizable patterns. For example, many of the measurable things in our world (e.g., IQ scores) follow a normal or Gaussian distribution.

When data are observed to follow probability distributions, and the data were collected in a way that maximizes the chances that they are representative of the larger population from which they were taken, it is possible to make estimates about the properties of that population.

Making inferences about some population by matching a sample of data with a probability distribution and estimating its parameters requires that the sample meet certain criteria. Specifically, you should have reason to believe that:

- the data follow the chosen type of distribution, and
- each point is independent and identically distributed (IID).

That is, the data were collected in a way that makes it likely that each data point is statistically independent from every other one, and each was drawn from a distribution with identical parameters. The nuances of how to collect data in ways that maximize the chances that your dataset satisfy these criteria are referred to as sampling techniques and fall outside the scope of this course.

Assuming you have a dataset that can be reasonably parameterized with a probability distribution, the process of estimating those parameters, also known as statistical inference is relatively straightforward.