Course › Probability: Introduction › Introduction to Probability › The Big Picture: Probability

🔖 Bookmark this page

The Big Picture: Probability

Recall the Big Picture—the four-step process that encompasses statistics (as it is presented in this course):
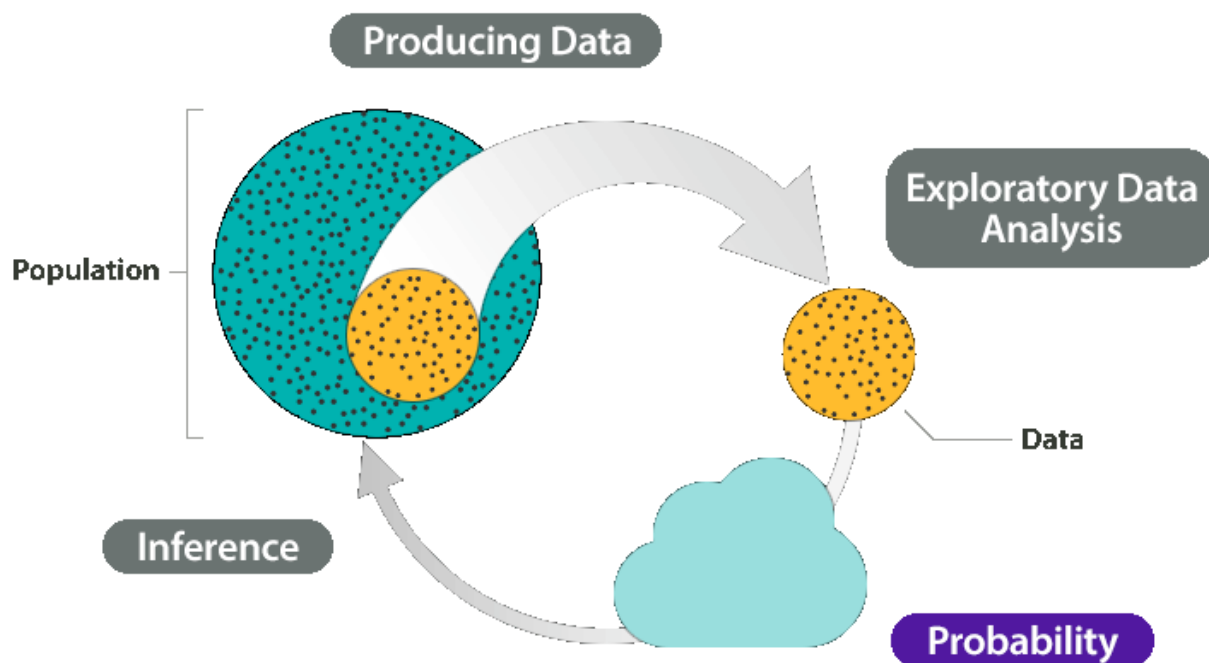
So far, we've discussed the first two steps:

Producing data—how data are obtained and what considerations affect the data production process.

Exploratory data analysis—tools that help us get a first feel for the data, by exposing their features using graphs and numbers.

(Recall that the structure of this course is such that exploratory data analysis was covered first, followed by producing data.)

Our eventual goal is **inference**—drawing reliable conclusions about the population based on what we've discovered in our sample. In order to really understand how inference works, though, we first need to talk about **probability**, because it is the underlying foundation for the methods of statistical inference. We use an example to explain why probability is so essential to inference.

First, here is the general idea: As we all know, the way statistics works is that we use a sample to learn about the population from which it was drawn. Ideally, the sample should be random so that it represents the population well.

Recall from the Sampling section that when we say a random sample represents the population *well*, we mean that there is no inherent bias in this sampling technique. It is important to acknowledge, though, that this does not mean that all random samples are necessarily "perfect." Random samples are still random, and therefore no random sample will be exactly the same as another. One random sample may give a fairly accurate representation of the population, while another random sample might be "off," purely due to chance. Unfortunately, when looking at a particular sample (which is what happens in practice), we will never know how much it differs from the population. This uncertainty is where probability comes into the picture. We use probability to quantify how much we expect random samples to vary. This gives us a way to draw conclusions about the population in the face of the uncertainty that is generated by the use of a random sample. The following example will illustrate this important point.

Example: Death Penalty

Suppose that we are interested in estimating the percentage of U.S. adults who favor the death penalty. In order to do so, we choose a random sample of 1,200 U.S. adults and ask their opinion: either in favor of or against the death penalty. We find that 744 out of the 1,200, or 62%, are in favor. (Comment: although this is only an example, this figure of 62% is quite realistic, given some recent polls). Here is a picture that illustrates what we have done and found in our example:
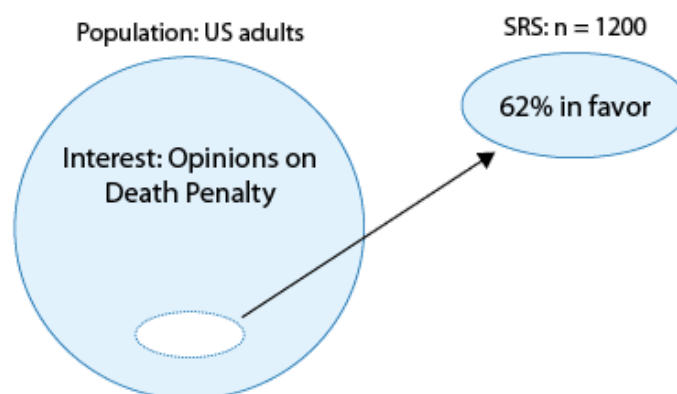
Our goal here is to do inference—learn and draw conclusions about the opinions of the entire population of U.S. adults regarding the death penalty, based on the opinions of only 1,200 of them.

Can we conclude that 62% of the population favors the death penalty? Another random sample could give a very different result. So we are uncertain. But since our sample is random, we know that our uncertainty is due to chance, and not due to problems with how the sample was collected. So we can use probability to describe the likelihood that our sample is within a desired level of accuracy. For example, probability can answer the question, "How likely is it that our sample estimate is no more than 3% from the true percentage of all U.S. adults who are in favor of the death penalty?"

The answer to this question (which we find using probability) is obviously going to have an important impact on the confidence we can attach to the inference step. In particular, if we find it quite unlikely that the sample percentage will be very different from the population percentage, then we have a lot of confidence that we can draw conclusions about the population based on the sample.