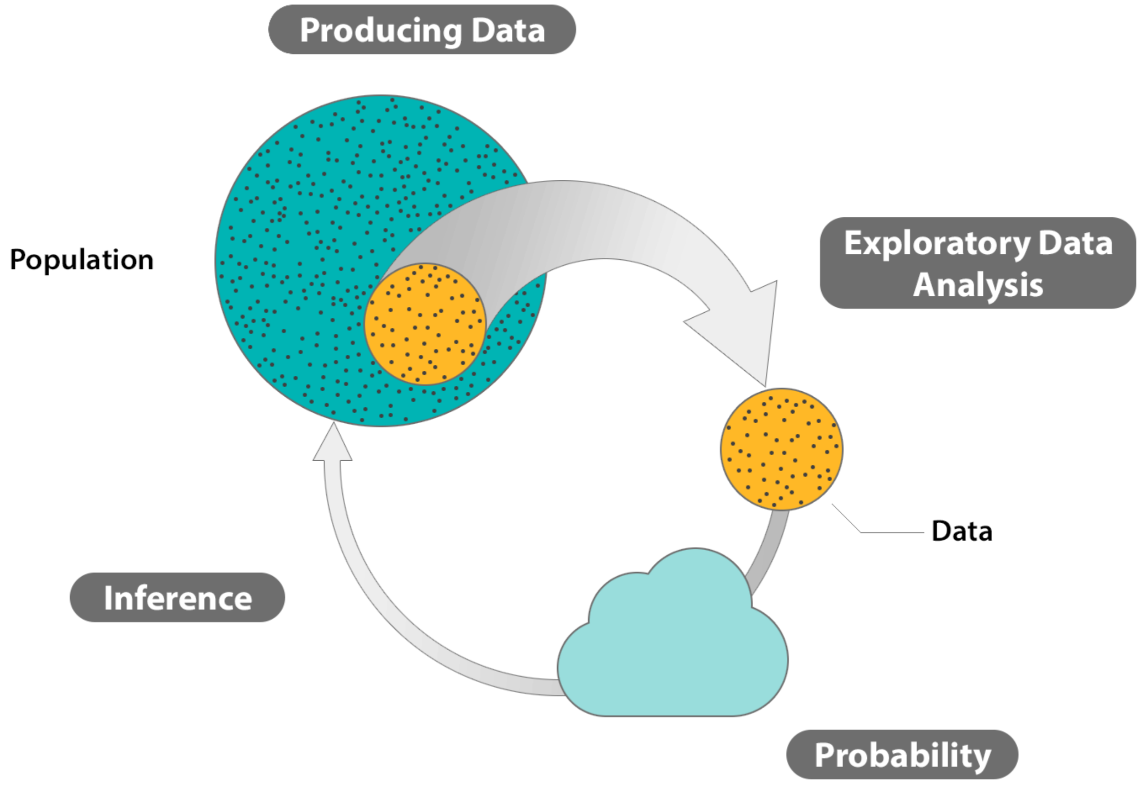
The Big Picture: Inference

Recall again the Big Picture, the four-step process that encompasses statistics: data production, exploratory data analysis, probability, and inference.

We are about to start the fourth part of the process and the final section of this course, where we draw on principles learned in the other units (exploratory data analysis, producing data, and probability) in order to accomplish what has been our ultimate goal all along: use a sample to infer (or draw conclusions) about the population from which it was drawn. The specific form of inference called for depends on the type of variables involved—either a single categorical or quantitative variable, or a combination of two variables whose relationship is of interest.



The purpose of this introduction is to review how we got here and how the previous sections fit together to allow us to make reliable inferences. Also, we will introduce the various forms of statistical inference that will be discussed in this section, and give a general outline of how this section is organized.

In the **Exploratory Data Analysis** sections, we learned to display and summarize data that were obtained from a sample. Regardless of whether we had one variable and we examined its distribution, or whether we had two variables and we examined the relationship between them, it was always understood that these summaries applied *only* to the data at hand; we did not attempt to make claims about the larger population from which the data were obtained.

Such generalizations were, however, a long-term goal from the very beginning of the course. For this reason, in the **Producing Data** sections, we took care to establish principles of sampling and study design that would be essential in order for us to claim that, to some extent, what is true for the sample should be also true for the larger population from which the sample originated. These principles should be kept in mind throughout this section on statistical inference, since the results that we will obtain will not hold if there was bias in the sampling process, or flaws in the study design under which variables' values were measured.

Perhaps the most important principle stressed in the Producing Data unit was that of randomization. Randomization is essential not only because it prevents bias but also because it permits us to rely on the laws of probability, which is the scientific study of random behavior.

In the **Probability** sections**,** we established basic laws for the behavior of random variables. We ultimately focused on two random variables of particular relevance: the sample mean ($\overline{X}$) and the sample proportion ($\hat{p}$), and the last module of the Probability unit was devoted to exploring their sampling distributions. We learned what probability theory tells us to expect from the values of the sample mean and the sample proportion, given that the corresponding population parameters—the population mean (μ) and the population proportion (p)—are known.

As we mentioned in that section, the value of such results is more theoretical than practical, since in real-life situations we seldom know what is true for the entire population. All we know is what we see in the sample, and we want to use this information to say something concrete about the larger population. Probability theory has set the stage to accomplish this: learning what to expect from the value of sample mean, given that population mean takes a certain value, teaches us (as we'll soon learn) what to expect from the value of the unknown population mean, given that a particular value of sample mean has been observed. Similarly, since we have established how sample proportion behaves relative to population proportion, we will now be able to turn this around and say something about the value of population proportion, based on an observed sample proportion. This process—inferring something about the population based on what is measured in the sample—is (as you know) called **statistical inference**.
