There are a variety of strategies that one can employ to setup a formal framework for making inferential statements. Often, there is literally a population of units (e.g. people, penguins, etc.) about which you want to make statements. In those cases it's clear where the uncertainty comes from (sampling from the population) and what exactly it is you're trying to estimate (some feature of the population). However, in other applications it might not be so clear what exactly is the population and what exactly it is you're trying to estimate. In those cases, you'll have to be more explicity about defining the population because there may be more than one possibility.

Time Series

Some processes are measured over time (every minute, every day, etc.). For example, we may be interested in analyzing data consisting of Apple's daily closing stock price for calendar year 2014. If we wanted to make an inference from this dataset, what would the population be? There are a few possibilities.

1. We might argue that the year 2014 was randomly sampled from the population of *all possible years* of data, so that inferences that we make apply to other years of the stock price.
2. We might say the Apple's stock represents a sample from the *entire stock market*, so that we can make inference about *other stocks* from this dataset.

Regardless of what you choose, it's important to make clear what population you are referring to before you attempt to make inference from the data.

Natural Processes

Natural phenomena, such as earthquakes, fires, hurricanes, weather-related phenomena, and other events that occur in nature, often are recorded over time and space. For purely temporal measurements, we might define the population in the same way that we defined the population above with the time series example. However, we may have data that is only measured in space. For example, we may have a map of the epicenters of all earthquakes that have occurred in an area. Then what is the population? One common approach is to say that there is an *unobserved stochastic process* that randomly drops earthquakes on to the area and that our data represent a random sample from this process. In that case, we are using the data to attempt to learn more about this unobserved process.

Data As Population

One technique that is always possible, but not commonly used, is to treat the dataset as a population. In this case, there is no inference because there's no sampling. Because your dataset *is* the population, there's no uncertainty about any characteristic of the population. This may not sound like a useful strategy but there are circumstances where it can be used to answer important questions. In particular, there are times where we do not care about things outside the dataset.

For example, it is common in organizations to analyze salary data to make sure that women are not being paid less than men for comparable work or that there are not major imbalances between employees of different ethnic

groups. In this setting, differences in salaries between different groups can be calculated in the dataset and one can see if the differences are large enough to be of concern. The point is that the data directly answer a question of interest, which is "Are there large salarly differences that need to be addressed?" In this case there's no need to make an inference about employees outside the organization (there are none, by definition) or to employees at other organizations over which you would not have any control. The dataset is the population and answers to any question regarding the population are in that dataset.

NOTE: Part of this reading was taken from *The Art of Data Science* by Peng & Matsui.