Representative Samples

4 min

Great! You've cleaned the data, decided what to do about missing data points, resolved any accuracy issues. You've made sure that all the questions can be answered by the data we have and gotten all stakeholders to agree on the research questions.

You are ready to do your analysis. Then you notice that all of the records are from New York State. You've been hired to work on the census for all of the North Atlantic. That includes multiple states in the U.S. and many regions of Canada. Where is the rest of the data?

You go back to the dataset and start to think about when it was collected. Right, Spring of 2020 - the border was closed. Census takers collected data in the region that they could: the areas that were convenient. This is a **convenience sample**. It's great for preliminary understanding, but not good for representing a broader population.

If we were to create a model to predict tree prettiness based on the variables in our dataset, it might only be relevant for trees in New York. We've introduced **bias** into our dataset by constraining our sample.

Convenience samples aren't the only type of sampling errors, but they are common. The goal of a sample is to represent a population. Any time a sample is made that does NOT reflect the entire population, it is a sampling error.

Best practice is to create a **sample** that represents the entire **population**.

The **population** is all of the trees in the North Atlantic region. The **sample** is the trees that we have data about (it will almost never be all of them).

The sample should look like the population in as many characteristics as possible. Therefore, our sample needs to include many different kinds of trees from many different locations.

There are a lot of techniques for creating representative samples, but they all have the same goal: to find a mix of observations that contains all of the features in the larger population.

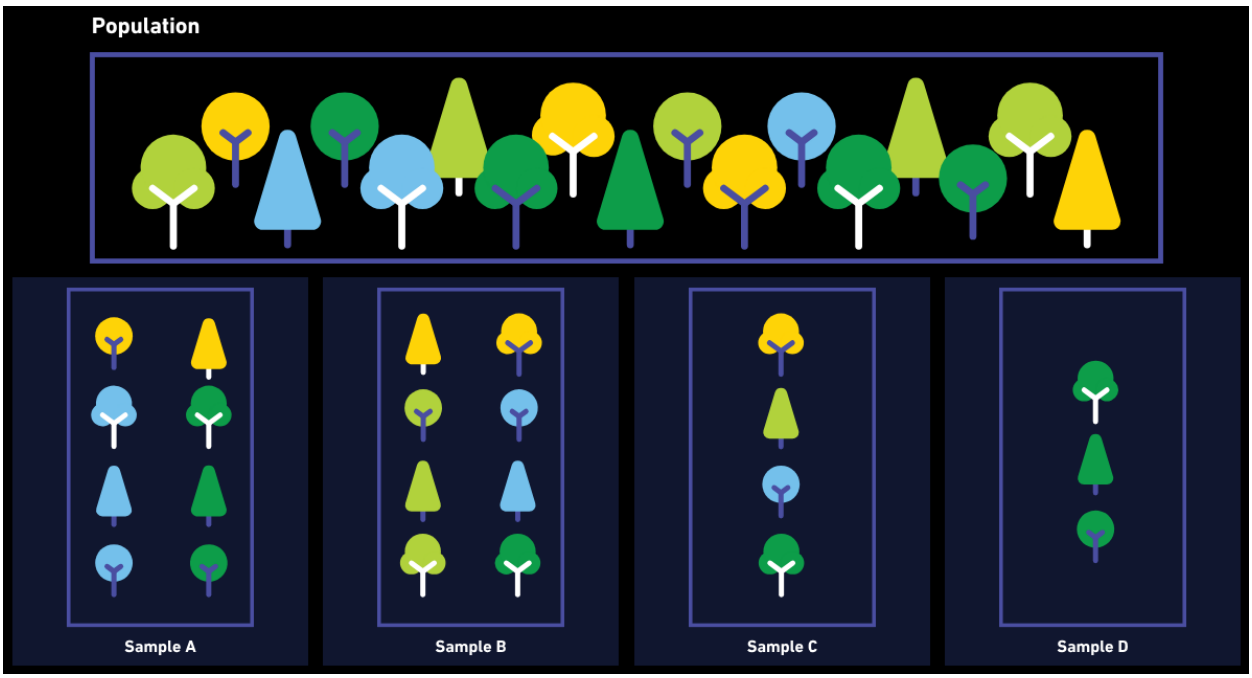
Instructions

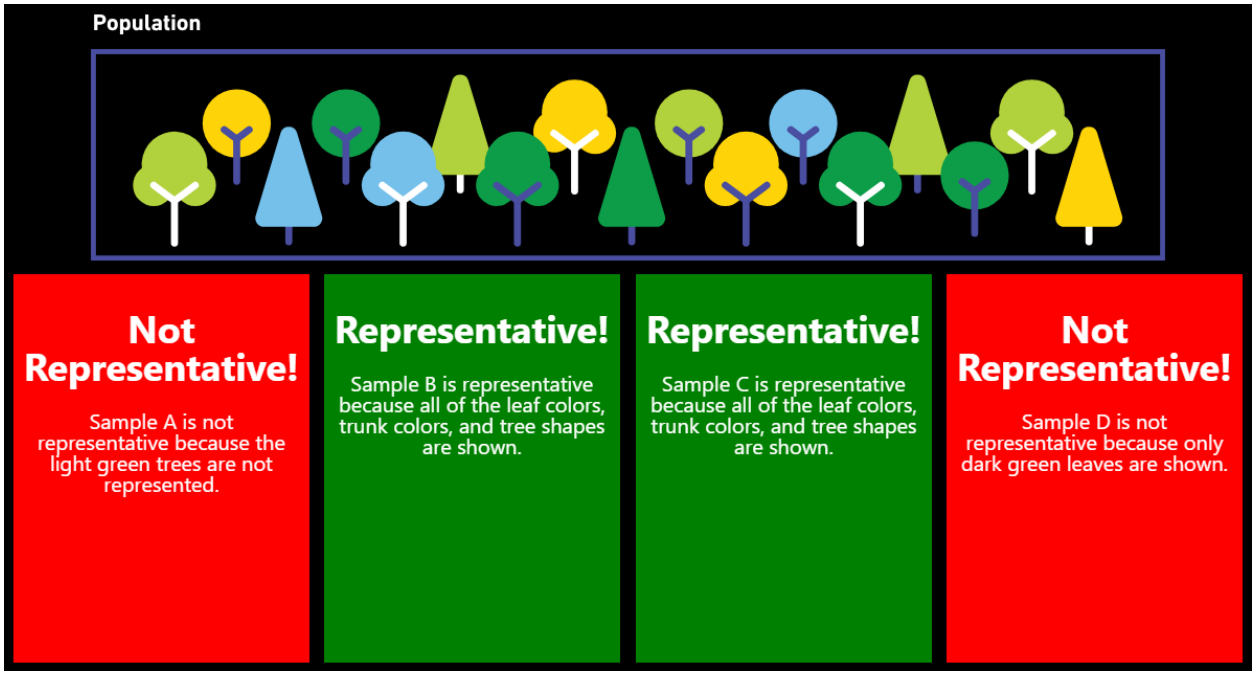
Review the Samples to the right. Every card has the same population of trees, but 4 different samples were chosen.

They are categorized on:

- Leaf color
- Trunk color
- Leaf type

Review each sample and decide if each is Representative or Non-Representative. Then click on the sample to check your answer.


Population
Sample A
Sample B
Sample C
Sample D


Population
Not Representative!
Sample A is not representative because the light green trees are not represented.
Representative!
Sample B is representative because all of the leaf colors, trunk colors, and tree shapes are shown.
Representative!
Sample C is representative because all of the leaf colors, trunk colors, and tree shapes are shown.
Not Representative!
Sample D is not representative because only dark green leaves are shown.