Validity

5 min

It's not just typos, mistakes, missing data, poor measurement, and duplicated observations that make a dataset low quality. We also have to make sure that our data actually measures what we think it is measuring. This is the **validity** of our dataset.

Validity is a special kind of quality measure because it's not just about the dataset, it's about the relationship between the dataset and its purpose. A dataset can be valid for one question and invalid for another.

Let's think again about our trees dataset. After we finished collecting the data, we thought of another question we wanted to answer: how old are our trees?

We know that you can measure the age of a tree by counting the rings, but we didn't do that. Let's say that we did measure the width of the tree.

We decide that since number of rings and width are related, we will use width as a proxy for the age. With that decision, we just compromised the **validity** of our dataset. Our data doesn't measure age, it measures width. And even though there is a relationship between the number of rings and the width, it's not a direct relationship and therefore cannot be substituted without affecting the validity of our dataset and measures.

Now let's say that we want to know how much our trees grow every year. We found a dataset for the same region from 20 years ago. We use the locations to match up the old and new measurements. But this data can tell us how much they grow every 20 years, not every year. If we try to use these two datasets to measure yearly growth, we will compromise the **validity** of the dataset again.

Using proxies and inappropriate time spans are just two ways to compromise the validity of a dataset. There are infinite ways in which a given dataset is not valid for answering a given question. The best way to spot issues with the validity of a dataset is to ask: Does this variable measure what I think it does?

Instructions

The quiz to the right has a list of research questions. Determine if each research question is valid and what variables are appropriate to address it using the dataset below.

Tree Census

ID	Height (ft)	Species	Location Type	Single	Distance (ft)	Prettiness
11246	5.90	Tulip	City	0	nan	3
11562	14.10	Red Oak	Highway	0	nan	5
11584	6.10	Honeylocust	Highway	0	nan	3
12034	26.30	Silver Maple	Undeveloped Area	nan	2.80	5
12139	10.40	Pin Oak	Undeveloped Area	nan	3.10	3
12438	5.40	Green Ash	Undeveloped Area	nan	nan	3
13007	7.50	nan	Highway	0	nan	3
13132	24.70	Silver Maple	Undeveloped Area	nan	nan	1
13280	nan	Red Oak	Highway	nan	nan	2
13281	nan	Honeylocust	City	1	3.20	5

Check your answers before moving to the next exercise.