Cross Sectional – It’s a snapshot of a group at a point in time

Longitudinal – Observes a group repeatedly over time

Cycle - From when to when a subset of data was collected. A dataset may comprise many cycles.

Population vs Sample — A population is a representation of all individuals of a specific type. Be it all individuals aged 11-20, all males between the ages of 40-45. In this case, it would be difficult, if not impossible to get data on all the individuals. What we can do however, is get a sample. A slice of the population.

There are different formulas for whether or not we have a population or a sample.

Use population formulas when either of these are true. You have data on the entire population, or if your sample is of a larger population, but you are not interested in applying your analysis to the entire population.

For example, if you have data on every student in your class, you can run statistical tests on it using either population formulae or sample formulae.

If you want to say for example “the variance of grades in this class” you would use Population Variance. However, if you wanted to use this variance as a representation of every student in every class, then you would have to use the population variance.

Cross sectional studies are supposed to be representative of the population. This means that every person in the population has an equal chance of randomly being selected. We should be selective in what our population is, and what our sample is.

If we are gathering data on the men aged 11-20 in the US then we would expect the distribution of the traits of people within the study, to match the population distribution of traits.

If the population of the US. is 57% non Latino White, for example, then if we are “truly” sampling a “random” slice of the population, then we should roughly get that.

A study done of men aged 11-20 in NYC for example, would have racial statistics that do not match the statistics of the US, as the race statistics in NYC are much different than that of the US. This is important to remember. We have to clearly define what our population is, and what our sample is, and we have to be able to determine whether we can truly make inferences about the population based on our sample.

It would be better to sample NYC randomly, and then make inferences about the population of men aged 11-20 NYC, than to try to use NYC and NYC only as a way to make inferences about the men aged 11-20 in the entirety of the US.

Sometimes data is deliberately oversampled. For example, if we were to sample black people and hispanic people more frequently than non hispanics and blacks, and it went against the population distribution for NYC, it could lead to skewed statistics. Sometimes, we want this. Maybe we are more interested in gathering data on these types of individuals, so that we can get a larger sample size of them than are represented in the population. We could want this so that we have a large enough sample size of people within that sample to accurately assess.

Our data is deliberately oversampled. This means that data from certain individuals appears at higher rates than they do in their representation in the US population. This makes it so that we have large enough groups of data about these individuals to make inferences about. It does however make our ability to generalize about the population a bit less accurate.

Libraries that you are going to be using

Pandas

NumPy

SciPy

StatsModels

Matplotlib

We are doing NSFG because the data is in a DAT format. We need to use Strata to more nicely create our DF.