SESSION 2: SAMPLES AND POPULATIONS

Session 2
Bias and Noise

Populations and Sample

- Population versus Sample: A population includes the universe of all instances of an object or phenomenon that you are trying to study. A sample is a subset of that population that you collect data on, and use, to make judgments about the behavior of the population.
- Time Series versus Cross Section: The data that you are trying to study can be a phenomenon that you observe over time (time series data) or across different subjects at a point in time (cross sectional data).
  - Time Series Example: If stock returns over time is your population, stock returns from 1960-2021 is a sample.
  - Cross sectional Example: If all publicly traded companies is your population, looking at only US companies or companies with market caps that exceed $10 million is a sample.

Why do we need sampling?

1. Practicality: If the population is too large to collect data on, and/or a subset (small or large) of the population is inaccessible, you have no choice but to sample the data.
2. Costs: Even if you could collect data on the entire population, the costs (in time and money) may outweigh the benefits of doing so.
3. Time trade off: Related to the second point is the question of how frequently you want to update the data. It is easier to update sampled data than data on the entire population.

Sampling Approaches

- Probability versus Non-probability Sampling: In a probability-based sample, the observations/subjects are picked at random. In a non-probability sample, the researcher hand picks the sample, based upon criteria that he or she picks.
- Variants of Random Sampling:
  - *Simple Random*: In a simple random sample, you pick your sample randomly across the entire population.
  - *Stratified Random*: In a stratified random sample, you first break your population down into groupings, and then randomly pick from within each of these groupings.
  - *Cluster Random*: In a cluster random sample, you break the population into groups, and then randomly select some of these groups and collect data on each group member.

Sampling Bias

- To the extent that you intend to extrapolate your findings on a sample to the entire population, you want to ensure that you don't have a biased sample.
- A biased sample is one that diverges from the population in its characteristics. That bias can arise for many reasons including:
  - Exclusion: Some parts of the population may not even make into the sampling universe.
  - Self-selection: Some parts of the population may be more easily accessible than other parts, given how you collect data.
  - Non-response: Some parts of the population may be less likely to respond to requests for data.
  - Survivorship: Success (or failure, sometimes) may make an observation more likely to be sampled.

And its consequences…

- If there is sampling bias, and you are or choose to be *unaware of that bias*, the extrapolations that you make from your sampling findings to the population will be biased.
- If there is sampling bias, and *you are aware of it, you can try to correct for it*, as you extrapolate your findings to the population.
- If *there is sampling bias, and you cannot correct for it*, you can narrow your findings to reflect the portion of the population that is represented by your sample.

Sampling Noise/ Error

- If a sample is unbiased, the results that you get from that sample can be extrapolated to the population, but with error (usually measured with a standard error on your forecast).
- That error is called sampling noise and will decrease as the sample size increases.
  - In a classic example, consider the odds of getting a head or a tail on an (unbiased) coin toss. Even though we know the population odds (50/50), you can get results that diverge on a small number of tosses. As the number of tosses increase, the sample numbers will also approach 50/50.
  - More generally, sampling noise is part and parcel of the process and all you can do is be transparent about it, and report it.

Independence + ID: Not just buzz words

- In almost any discussion of sampling and statistics, you will hear the words "*independence*" and "*identical distributions*" thrown in as pre-requisites or at least good qualities in a sample.
  - Independence: Events are independent when whether an event occurs or not is not determined by other events occurring.
    - Coin tosses are a classic example of independence
    - Are stock price changes independent?
  - Identical Distributions: Each event draws from the same probability distribution.
    - Coin tosses draw from the same distribution (50/50)
    - Do stock price changes draw from the same distribution?
- In finance, researchers often assume independence and identical distributions, in making assertions based upon samples, but the truth is that both characteristics are hard to find.

The Law of Large Numbers

- The Law of Large Numbers: As the number of observations in a sample increases, the sample average will approach the population average (or the true average), if the observations are independent(I) and identically distributed (ID).
- There are a few exceptions to the law of large numbers.
  - The first is if the sample observations are not independent or identically distributed.
  - The other is for distributions with fat tails (higher chance of extreme outcomes).
- The Weak Law of Large Numbers: Even when observations are not independent or identically distributed, the sample average will approach the population average, albeit slower and only if the variance is finite.

Sampling Error Theorems/Propositions

- Central Limit Theorem: Even if your population distribution is non-normal, the means of any samples that you pull from that distribution will follow a normal distribution, as long as you the observations are independent and identically distributed.
- Chebyshev's inequality : In a normal distribution, for instance, roughly two thirds of the values have to fall within one standard deviation of the mean and 95% within two standard deviations. Chebyshev's inequality is more general, stating that a minimum of just 75% of values must lie within two standard deviations of the mean and 88.89% within three standard deviations for a broad range of different probability distributions.