Topic #4: Sample and Population

In statistics, a sample is a subset of a population. Typically, the population is very large, making a census or a complete enumeration of all the values in the population impractical or impossible. The sample represents a subset of manageable size. Samples are collected and statistics are calculated from the samples so that one can make inferences or extrapolations from the sample to the population. This process of collecting information from a sample is referred to as sampling.

Samples are expected to be selected in such a way as to avoid presenting a biased view of the population. The sample will be unrepresentative of the population if certain members of the population are excluded from any possible sample. For example, if a researcher is interested in the drug usage patterns among teenagers, but collects the sample from the local schools, the sample is biased because it excludes drop-outs and home-schooled teenagers. Biases can also occur if some members of the population are more likely or less likely to be included in the sample than other members of the population. So the sample collected from schools is also biased because students who miss a lot of school days because of a chronic illness will be less likely to be selected for a sample than students who attend very regularly.

The best way to avoid a biased or unrepresentative sample is to select a random sample, also known as a probability sample. A random sample is defined as a sample where the probability that any individual member from the population being selected as part of the sample is exactly the same as any other individual member of the population. Several types of random samples are simple random samples, systematic samples, stratified random samples, and cluster random samples.

A sample that is not random is called a nonrandom sample or a nonprobability sample. Some examples of nonrandom samples are

convenience samples, judgment samples, purposive samples, quota samples, and snowball samples.

In mathematical terms, given a random variable X with distribution F, a sample of length n\in\mathbb{N} is a set of n independent, identically distributed (iid) random variables with distribution F. It concretely represents n experiments in which we measure the same quantity. For example, if X represents the height of an individual and we measure n individuals, Xi will be the height of the i-th individual. Note that a sample of random variables (i.e. a set of measurable functions) must not be confused with the realisations of these variables (which are the values that these random variables take). In other words, Xi is a function representing the mesure at the i-th experiment and xi = Xi(?) is the value we actually get when making the measure.