1 Introduction to Statistics

Having developed a sufficient foundational probability background. We are now ready to apply it to study statistics. We first start by defining a simple data model to analyze.

1.1 Random Samples

Data collected in an experiment represents several observations on a variable of interest. Our focus will be on one type of collecting data called random sampling.

Definition 1 (Random Samples)**.** The random variables $X_1, \ldots, X_n$ are called a random sample of size n from the population F if $X_1, \ldots, X_n$ independent and identically distributed random variables with CDF F.

Our goal is to study facts about the population F using many realizations of it. Random sampling involve independent trials of the same experiment. To simplify our problem, suppose that we know the general behavior of the population, but we are unsure of precise behavior encoded by a parameter θ. The goal is to learn θ from our samples.

1.1.1 Types of Random Samples

If you do not want to ask the same person twice for a survey, so sampling removes someone in a finite population, and hence changes the remaining population for the next sample. This is sampling without replacement, and it captures the fact that are independent, but not identically distributed.

Definition 2 (Sampling without replacement)**.** Sampling without replacement follows the following rule. Let $\{x_1, \ldots, x_N\}$ and suppose we select one element from that set uniformly at random. Let X_1 denote the value of this first sample, so $X_1 = x$. The random element X_1 is removed from the set, and we sample X_2 from the remaining uniformly at random, and so on.

A less realistic way of sampling is simple random sample, which is sampling with replacement. This makes the model easier, but it is less realistic.

Definition 3 (Sampling with replacement)**.** Sampling with replacement follows the following rule. Let $\{x_1, \ldots, x_N\}$ and suppose we select one element from that set uniformly at random. Let X_1 denote the value of this first sample, so $X_1 = x$. The random element X_1 is put back in the set, and we sample X_2 from the remaining uniformly at random, and so on.

Simple random sampling produces a random sample where F is uniform on $\{x_1, \ldots, x_N\}$. The reason we call random sampling as sampling from an infinite population is because in the limit, both types of sampling are equivalent.

Proposition 1

Consider a population of N elements. Suppose that we want to sample K elements. Let $(X_1, \ldots, X_K)$ and $(Y_1, \ldots, Y_K)$ denote samples obtained via sampling without replacement and sampling with replacement respectively. We have

$$(X_1, \ldots, X_K) \to (Y_1, \ldots, Y_K)$$

in distribution as $N \to \infty$.

Throughout the rest of the course, we will work with large populations, so we will work with random samples, without losing so much precision.

1.2 Summary Statistics

We are often not interested in the entire sample, but a summary of the sample called a statistic.

Definition 4 (Statistic)**.** Let $X_1, \dots, X_n$ be a random sample of size n from a population and let $T(x_1, \dots, x_n)$ be a real-valued (non-random) function (or vector). The random variable $Y = T(X_1, \dots, X_n)$ is called the *statistic* and the distribution of Y is called the *sampling distribution*.

In statistics, our main goal is often to understand the *sampling statistic* because it will tell us if the summary is expected or rare. The function T can be general, but it usually never depends on the parameters defining our model.

1.2.1 Examples of Common Statistics

The moments were the idealized summary statics for random variables. hence, two common statistics measure estimate the mean and variance of population.

Definition 5 (Sample Mean)**.** The sample mean is the average of the sample

$$\bar{X} = \sum_{i=1}^{n} X_i$$

By the central limit theorem, we approximately know the sampling distribution for every sample.

Definition 6 (Sample Mean)**.** The sample variance is the variance of the sample

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

The sample standard deviation is S.

It might be somewhat unexpected that we are normalizing by $\frac{1}{n-1}$ instead of n.

1.3 Example Problems

Problem 1.1. Let $X_1, \dots, X_N$ be samples from a population $\{x_1, \dots, x_N\}$ without replacement. What is the marginal distribution of X_i?

Solution 1.1. We have that

$$\mathbb{P}(X_1 = x) = \frac{1}{N}$$

for each $x \in \{x_1, \dots, x_N\}$. Since the second sample depends on the outcome of the first, we need to condition on the outcome fo the first draw

$$\mathbb{P}(X_2 = x) = \sum_{i=1}^{N} \mathbb{P}(X_2 = x, X_1 = x_i) = \sum_{i=1}^{N} \mathbb{P}(X_2 = x \mid X_1 = x_i)\, \mathbb{P}(X_1 = x_i).$$

If $X_1 = x$, then $\mathbb{P}(X_2 = x \mid X_1 = x_i) = 0$, but for all others it is $\frac{1}{N-1}$. Therefore,

$$\mathbb{P}(X_2 = x) = (N-1)\frac{1}{N-1}\frac{1}{N} = \frac{1}{N}.$$

A similar computation shows that

$$\begin{aligned}\mathbb{P}(X_3 = x) &= \sum_{j=1}^{N}\sum_{i=1}^{N} \mathbb{P}(X_3 = x, X_2 = x_j, X_1 = x_i) \\ &= \sum_{j=1}^{N}\sum_{i=1}^{N} \mathbb{P}(X_3 = x | X_2 = x_j, X_1 = x_i)\,\mathbb{P}(X_2 = x_j | X_1 = x_i)\,\mathbb{P}(X_1 = x_i) \\ &= \frac{N-2}{N-2}\frac{N-1}{N-1}\frac{1}{N} = \frac{1}{N},\end{aligned}$$

and so on.

Alternative Solution: Instead of looking at the problem sequentially, we can look at the problem as a whole. We can denote the outcome of a random sample as a uniform permutation of $(x_1, \ldots, x_N)$. The order of the permutation determines the outcomes of $X_1, X_2, \ldots, X_N$. This is clear by symmetry. Then for a fixed value $x \in (x_1, \ldots, x_N)$, we have

$$\mathbb{P}(X_j = x) = \frac{1}{N}$$

since x can appear in any of the N spots with the same probability.

1.4 Proofs of Key Results