8.1.1 Random Sampling

When collecting data, we often make several observations on a random variable. For example, suppose that our goal is to investigate the height distribution of people in a well defined population (i.e., adults between 25 and 50 in a certain country). To do this, we define random variables X_1, X_2, X_3, $\ldots$, X_n as follows: We choose a random sample of size n with replacement from the population and let X_i be the height of the i th chosen person. More specifically,

1. We chose a person uniformly at random from the population and let X_1 be the height of that person. Here, every person in the population has the same chance of being chosen.
2. To determine the value of X_2, again we choose a person uniformly (and independently from the first person) at random and let X_2 be the height of that person. Again, every person in the population has the same chance of being chosen.
3. In general, X_i is the height of the ith person that is chosen uniformly and independently from the population.

You might ask why do we do the sampling with replacement? In practice, we often do the sampling without replacement, that is, we do not allow one person to be chosen twice. However, if the population is large, then the probability of choosing one person twice is extremely low, and it can be shown that the results obtained from sampling with replacement are very close to the results obtained using sampling without replacement. The big advantage of sampling with replacement (the above procedure) is that X_i's will be independent and this makes the analysis much simpler.

Now for example, if we would like to estimate the average height in the population, we may define an estimator as

$$\hat{\Theta} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

The random variables X_1, X_2, X_3, $\ldots$, X_n defined above are independent and identically distributed (i.i.d.) and we refer to them collectively as a (simple) random sample.

The collection of random variables X_1, X_2, X_3, $\ldots$, X_n is said to be a **random sample** of size n if they are independent and identically distributed (i.i.d.), i.e.,

1. X_1, X_2, X_3, $\ldots$, X_n are independent random variables, and
2. they have the same distribution, i.e,

$$F_{X_1}(x) = F_{X_2}(x) = \ldots = F_{X_n}(x), \qquad \text{for all } x \in \mathbb{R}.$$

In the above example, the random variable $\hat{\Theta} = \frac{X_1+X_2+\cdots+X_n}{n}$ is called a **point estimator** for the average height in the population. After performing the above experiment, we will obtain $\hat{\Theta} = \hat{\theta}$. Here, $\hat{\theta}$ is called an **estimate** of the average height in the population. In general, a point estimator is a function of the random sample $\hat{\Theta} = h(X_1, X_2, \cdots, X_n)$ that is used to estimate an unknown quantity.

It is worth noting that there are different methods for sampling from a population. We refer to the above sampling method as *simple random sampling*. In general, "sampling is concerned with the selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population" [18]. Nevertheless, for the material that we cover in this book simple random sampling is sufficient. Unless otherwise stated, when we refer to random samples, we assume they are simple random samples.

Some Properties of Random Samples:

Since we will be working with random samples, we would like to review some properties of random samples in this section. Here, we assume that X_1, X_2, X_3, $\ldots$, X_n are a random sample. Specifically, we assume

1. the X_i's are independent;
2. $F_{X_1}(x) = F_{X_2}(x) = \ldots = F_{X_n}(x) = F_X(x)$;
3. $EX_i = EX = \mu < \infty$;
4. $0 < \text{Var}(X_i) = \text{Var}(X) = \sigma^2 < \infty$.

Sample Mean:

The sample mean is defined as

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

Another common notation for the sample mean is M_n. Since X_i are assumed to have the CDF $F_X(x)$, the sample mean is sometimes denoted by $M_n(X)$ to indicate the distribution of X_i's.

Properties of the sample mean

1. $E\overline{X} = \mu$.
2. $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$.
3. Weak Law of Large Numbers (WLLN):

$$\lim_{n \to \infty} P(|\overline{X} - \mu| \geq \epsilon) = 0.$$

4. Central Limit Theorem: The random variable

$$Z_n = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = \frac{X_1 + X_2 + \ldots + X_n - n\mu}{\sqrt{n}\sigma}$$

converges in distribution to the standard normal random variable as n goes to infinity, that is

$$\lim_{n \to \infty} P(Z_n \leq x) = \Phi(x), \qquad \text{for all } x \in \mathbb{R}$$

where $\Phi(x)$ is the standard normal CDF.

Order Statistics:

Given a random sample, we might be interested in quantities such as the largest, the smallest, or the middle value in the sample. Thus, we often order the observed data from the smallest to the largest. We call the resulting ordered random variables *order statistics*. More specifically, let X_1, X_2, X_3, $\ldots$, X_n be a random sample from a continuous distribution with CDF $F_X(x)$. Let us order X_i's from the smallest to the largest and denote the resulting sequence of random variables as

$$X_{(1)}, X_{(2)}, \cdots, X_{(n)}.$$

Thus, we have

$$X_{(1)} = \min\left(X_1, X_2, \cdots, X_n\right);$$

and

$$X_{(n)} = \max\left(X_1, X_2, \cdots, X_n\right).$$

We call $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ the **order statistics** of the random sample X_1, X_2, X_3, ..., X_n. We are often interested in the PDFs or CDFs of the $X_{(i)}$'s. The following theorem provides these functions.

Theorem 8.1

Let X_1, X_2, ..., X_n be a random sample from a continuous distribution with CDF $F_X(x)$ and PDF $f_X(x)$. Let $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ be the order statistics of X_1, X_2, X_3, ..., X_n. Then the CDF and PDF of $X_{(i)}$ are given by

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} f_X(x)\big[F_X(x)\big]^{i-1}\big[1 - F_X(x)\big]^{n-i},$$

$$F_{X_{(i)}}(x) = \sum_{k=i}^{n} \binom{n}{k} \big[F_X(x)\big]^{k}\big[1 - F_X(x)\big]^{n-k}.$$

Also, the joint PDF of $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ is given by

$$f_{X_{(1)},\cdots,X_{(n)}}(x_1, x_2, \cdots, x_n) = \begin{cases} n! f_X(x_1) f_X(x_2) \cdots f_X(x_n) & \text{for } x_1 \leq x_2 \leq x_2 \cdots \leq x_n \\ \\ 0 & \text{otherwise} \end{cases}$$

A method to prove the above theorem is outlined in the End of Chapter Problems section. Let's look at an example.

Example 8.1

Let X_1, X_2, X_3, X_4 be a random sample from the $Uniform(0,1)$ distribution, and let $X_{(1)}$, $X_{(2)}$, $X_{(3)}$, $X_{(4)}$. Find the PDFs of $X_{(1)}$, $X_{(2)}$, and $X_{(4)}$.

Solution

Here, the ranges of the random variables are $[0,1]$, so the PDFs and CDFs are zero outside of $[0,1]$. We have

$$f_X(x)=1, \qquad \text{for } x\in[0,1],$$

and

$$F_X(x)=x, \qquad \text{for } x\in[0,1].$$

By Theorem 8.1, we obtain

$$\begin{aligned} f_{X_{(1)}}(x) &= \frac{4!}{(1-1)!(4-1)!} f_X(x)\big[F_X(x)\big]^{1-1}\big[1-F_X(x)\big]^{4-1} \\ &= 4f_X(x)\big[1-F_X(x)\big]^3 \\ &= 4(1-x)^3, \qquad \text{for } x\in[0,1]. \end{aligned}$$

$$\begin{aligned} f_{X_{(2)}}(x) &= \frac{4!}{(2-1)!(4-2)!} f_X(x)\big[F_X(x)\big]^{2-1}\big[1-F_X(x)\big]^{4-2} \\ &= 12f_X(x)F_X(x)\big[1-F_X(x)\big]^2 \\ &= 12x(1-x)^2, \qquad \text{for } x\in[0,1]. \end{aligned}$$

$$\begin{aligned} f_{X_{(4)}}(x) &= \frac{4!}{(4-1)!(4-4)!} f_X(x)\big[F_X(x)\big]^{4-1}\big[1-F_X(x)\big]^{4-4} \\ &= 4f_X(x)\big[F_X(x)\big]^3 \\ &= 4x^3, \qquad \text{for } x\in[0,1]. \end{aligned}$$