5 Estimation - I

5.1 Random sample

To understand the concept of a random sample, we first need to understand the concept of a **population**. The population contains all the outcomes of a random variable and the distribution of the outcomes in the population is reflected by the probability distribution of the corresponding random variable X. A population may be finite or infinite.

A random sample is a sample drawn from the population such that the probability of each element in the population of being selected is the same and is given by the *pdf* (or *pmf*) of the random variable. Each selection in a sample of size n can be thought of as a random variable X_i with a *pdf* identical to that of the population random variable X. Hence, the probability of any element of the population being selected as the i^{th} element of the sample is the same as its probability of being in the population. For example if the random variable X reflects gender, and males and females are equally likely to be in the population, then the probability of a female (or male) being selected as the 1st (or 2nd, 3rd etc.) element of a random sample is 0.5.

Hence, a random sample is a collection of random variables, $X_1,\ X_2, ..., X_n$ that are mutually independent, with their *pdf*s being equal to each other and to the *pdf* of the random variable corresponding to the underlying population, i.e. $f_{X_i}(x) = f_X(x), i = 1$ to n (sometimes the subscript X is dropped for the population *pdf* and it is just referred to as $f(x)$). Such a collection of random variables is also called *independent and identically distributed* or *iid*.

An observation is the value that each collected random variable assumes (say $X_1 = x_1,\ X_2 = x_2, ...,\ X_n = x_n$). In the gender example, if the random variable denoted female as 1 and male as 0, then the observations would look like $X_1 = 1,\ X_2 = 0$, etc.

For a finite population, we often come across the question of sampling with and without replacement. When sampling without replacement, the independence assumption gets violated as the probability of an element getting selected as the second observation in the sample is clearly dependent on what the first selection was. However, when the population size is large, we can assume that the dependence is very weak and the selections are nearly independent.

5.2 Statistic

Let the random variables $X_1, X_2, ..., X_n$ be a random sample of size n from the population with a *pdf* $f(x)$. A *statistic*, sometimes also called a *sample statistic*, is any function of these random variables i.e. $T = g(X_1, X_2, .., X_n)$. Since the inputs to the function are random variables, the statistic itself is also a random variable.

5.2.1 Sample mean

The sample mean $\bar{X}$ is defined as

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

It is important to note that the sample mean is also a random variable. For example, think of rolling a six sided dice. The population consists of six outcomes all equally likely. Now, let us take a sample of two outcomes and calculate the sample mean. This sample mean can take the values 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5 and 6. We can also assign probabilities to each of these values using he probabilities of drawing the sample that will have this sample mean. For example, the only sample that will have a mean of 6 is 6,6. Hence, the probability of the sample mean being 6 is $\frac{1}{36}$.

In contrast, the population mean or expected value μ is a number. In this case, it is 3.5. [1]

5.2.2 Sample variance

The sample variance s^2 is defined as

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

5.3 Distribution of the sample mean

Since the sample mean is a random variable, we can ask what its expected value $E[\bar{X}]$ and its variance $Var(\bar{X})$ is. We can use the properties of expectation and variance seen earlier to show that

$$\begin{aligned}
E[\bar{X}] &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n}E\left[\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\
&= \frac{1}{n}nE[X] \qquad (\because E[X_i] = E[X]\ \forall i) \\
&= E[X] = \mu
\end{aligned}$$

Unsurprisingly, the expected value of the sample mean is the population mean.

[1] In many cases, we will not know the exact probability distribution of the population but will know the class of distributions (eg, uniform, normal etc) that it belongs to. In such cases the population mean (and other population moments) will be functions of some unknown distribution parameters. For example, think back to the Bernoulli distribution example used in Lecture note 3, where the expected value was equal to the unknown probability parameter p.

The variance of the sample mean is

$$\begin{aligned} Var(\bar{X}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X) \qquad (\because X_i \text{ are independent of each other}) \\ &= \frac{1}{n^2} n Var(X) = \frac{\sigma^2}{n} \end{aligned}$$

Hence, the variance of the sample mean is given by the population variance divided by the sample size.

5.3.1 Central limit theorem

The central limit theorem allows us to know not just the mean and variance of the sample mean, but also know something about its overall distribution. The theorem says that if a sample of size n is taken from a population with any distribution[2] with population mean μ and population variance σ^2, then as $n \to \infty$, the distribution of the sample mean approaches a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

In practice, if n is sufficiently large, we can assume that the distribution of $\bar{X}$ is normal without knowing the underlying population distribution. This is a powerful result.

5.4 The normal distribution

The normal distribution is a class of symmetric continuous probability distributions that is used to explain many naturally occurring distributions like heights

[2]Any distribution with a finite mean and variance

of individuals, weights of a certain species of animal etc. The *pdf* of a normally distributed random variable X is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The distribution is defined by two parameters μ and σ. These correspond to the mean and standard deviation respectively.

$$E[X] = \mu$$

$$Var(X) = \sigma^2$$

As the *pdf* is difficult to integrate, in practice we convert any normally distributed random variable to the standard normal distribution, which has mean 0 and variance 1. Let X be a normally distributed random variable with mean μ and variance σ^2. In short this is written as $X \sim N(\mu, \sigma^2)$. We convert this into the standard normal random variable Z as follows.

$$Z = \frac{X - \mu}{\sigma}$$

Z is distributed as $N(0, 1)$ and its *cdf* is made available as statistical tables, to be used for *estimation*.

5.5 Point estimation

Let us go back to the unfair coin example we saw in Lecture notes 3. An unfair coin has an unknown probability of turning up heads. This process is captured by the Bernoulli distribution with the following *pmf*.

$$f_X(x) = \begin{cases} p & for\ x = 1, \\ 1 - p & for\ x = 0, \\ 0 & for\ x \neq 0, 1 \end{cases}$$

We showed that the expected value of this distribution

$$E[X] = p$$

Now we can take a sample of n independent observations, i.e. we toss the coin n times, and find the sample mean. This sample mean can be used as an *estimator* for the population mean and hence for the parameter p.

An **estimator** is a sample statistic, i.e. it is a function of the random variables comprising the random sample, that is used to estimate the value of some population parameter θ .

Any function could be an estimator. The median of the sample, the average of the highest and the lowest observations, could all be estimators for the population mean.

There are some properties that we look for in an estimator.

1. **Unbiasedness:** An unbiased estimator is one whose expected value is equal to the parameter being estimated. Therefore, the sample mean, $\bar{X}$ is always an unbiased estimator of the population mean μ because $E[\bar{X} = \mu]$. Similarly, we can show that the sample variance is an unbiased estimator of the population variance.

2. **Efficiency:** One estimator is considered more efficient than another when the variance of the first is lower than the other. Consider a random sample of size n, and another of size m from the same population such that $m > n$. Let us denote the sample means from these two samples as $\bar{X}_n$ and $\bar{X}_m$ respectively. Both of these sample means are unbiased estimators of the population mean, but their variances are different.

$$Var(\bar{X}_n) = \frac{\sigma^2}{n} > \frac{\sigma^2}{m} = Var(\bar{X}_m)$$

Therefore, $\bar{X}_m$ is a more efficient estimator of the population mean than $\bar{X}_n$.

3. **Consistency:** An estimator is consistent if as the sample size increases, the estimator gets closer to the population parameter being estimated.

Proof that the sample variance is an unbiased estimator of the population variance (extra)

Let us define the random variables $Y_i = X_i - \mu$. Therefore, $E[Y_i] = 0$, $\bar{Y} = \bar{X} - \mu$, $Var(Y_i) = Var(X_i) = \sigma_X^2$, and $Var(\bar{Y}) = Var(\bar{X}) = \frac{\sigma_X^2}{n}$.

Now, the sample variance is given by

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}((X_i - \mu) - (\bar{X} - \mu))^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \\ &= \frac{1}{n-1}\left[\left(\sum_{i=1}^{n} Y_i^2\right) - n\bar{Y}^2\right] \end{aligned}$$

The expected value of the sample variance is

$$E[S^2] = \frac{1}{n-1}\left(E\left[\sum_{i=1}^{n} Y_i^2\right] - nE[\bar{Y}^2]\right)$$

Now,

$$\begin{aligned} E\left[\sum_{i=1}^{n} Y_i^2\right] &= \sum_{i=1}^{n} E\left[Y_i^2\right] \\ &= \sum_{i=1}^{n} E\left[(Y_i^2 - 0)\right] \\ &= \sum_{i=1}^{n} E\left[(Y_i^2 - E[Y_i])\right] \\ &= nVar(Y_i) = n\sigma_X^2 \end{aligned}$$

And,

$$\begin{aligned} nE[\bar{Y}^2] &= nE[(\bar{Y}-0)^2] \\ &= nE[(\bar{Y}-E[\bar{Y}])^2] \\ &= nVar(\bar{Y}) = n\frac{\sigma_X^2}{n} = \sigma_X^2 \end{aligned}$$

Therefore,

$$\begin{aligned} E[S^2] &= \frac{1}{n-1}\left(E\left[\sum_{i=1}^{n} Y_i^2\right] - nE[\bar{Y}^2]\right) \\ &= \frac{1}{n-1}(n\sigma_X^2 - \sigma_X^2) = \sigma_X^2 \end{aligned}$$

Hence, we have shown that the sample variance calculated with $n-1$ in the denominator is an unbiased estimator of the population variance.