Sampling

In formulating a problem, it is important to identify the relevant **population**, the entire group of items about which some information is desired.

We are interested in drawing inferences about several attributes of a population. But it is usually prohibitively expensive to study every single element of a population.

So we use a **sample** of the population in order to draw conclusions about the characteristics of the population. This is known as **statistical inference**.

Several types of sampling are possible. We focus only on **random sampling**.

How do we estimate these characteristics from the sample? How much confidence should we have in these estimates?

To answer such questions we need to turn to **sampling distributions**.

Sampling Distributions and Estimation

Let’s be clear between what comes from the sample and what comes from the population.

A **population parameter** (or **parameter**) is a characteristic of the population, it is fixed, but unknown.

A **sample statistic** (or **statistic**) is a function of the observed values of random variables that does not contain any unknown parameters (eg. the sample mean or sample variance).

An **estimator** is a sample statistic that we will use to estimate a population parameter.

An **estimate** is the specific value of the estimator.

Sample statistics are random variables. If we draw a different sample of the same size, we will get a different value for say the mean. If we repeat this process, we will get a large number of values for the mean. This is called **sampling variation**.

We can generate a **sampling distribution** for this mean. This is the probability distribution that describes the population of all possible values of the sample mean.

Let’s think about this more carefully.

Sampling distributions of the sample mean ($\bar{X}$) and sample variance (s^2) are of significant interest in econometrics, especially when the population is normal.

Suppose X is a random variable that has a normal distribution with mean μ and variance σ^2. In other words, $X \sim N(\mu, \sigma^2)$.

Let's draw a sample of size n from the population: $X_1, \dots, X_n$.
This sample is **iid** (identically and independently distributed).

The sample mean and sample variance are defined as:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

What is the sampling distribution of $\bar{X}$?

Sampling from a normal population

The sample mean is just a linear combination of n random normal variables. A linear combination also has a normal distribution. Further, it can be shown that $\bar{X}$ has a mean of μ and $Var(\bar{X}) = \sigma^2/n$. Thus, $\bar{X} \sim N(\mu, \sigma^2/n)$.

The distribution of $Z = \frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}} \sim N(0, 1)$.

A sample statistic is said to be an **unbiased estimator** a parameter if the mean of the sampling distribution of the statistic is equal to the value of the parameter.

Is $\bar{X}$ an unbiased estimator of μ?

The standard deviation of the sampling distribution of $\bar{X}$ is $/\sqrt{n}$. How do we compute this if we do not know σ? We can estimate it using s, the standard deviation of the sample, where $s = \sqrt{s^2}$.

An estimate of the standard deviation is called a **standard error**

$$\text{standard error } \bar{X} = s/\sqrt{n}.$$

Large-Sample Distributions

Two very useful properties when the sample size is large:

- **Law of large numbers** (LLN)
  - As the sample size n increases, the sample mean of a set of random variables approaches its expected value

- **Central limit theorem** (CLT)
  - Let $X_1, \ldots, X_n$ be a random sample from the same distribution with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Then the sampling distribution of the random variable $Z_n = (\bar{X} - \mu)/[\sigma/\sqrt{n}]$ converges to the standard normal $N(0, 1)$ as n converges to infinity.