Lesson 1: Sampling, Sampling Error, and the Distribution of the Sample Mean

SAMPLING AND ESTIMATION

In a **simple random sample**, each member of the population has the same probability or likelihood of being included in the sample.

Example

Assume that our population consists of 10 balls labeled with numbers 1 to 10.

- Drawing a *random* sample of 3 balls from this population of 10 balls would require that:
  - Each ball has an equal chance of being chosen in the sample
  - Each combination of balls has an identical chance of being the chosen sample as any other combination

In **systematic sampling**, every *k*th member in the population list is selected until the desired sample size is reached.

Sampling error is the error caused by observing a sample instead of the entire population to draw conclusions relating to population parameters. It equals the difference between a sample statistic and the corresponding population parameter.

A **sampling distribution** is the probability distribution of a given sample statistic under repeated sampling of the population.

Example

A random sample of 50 stocks is selected from a population of 10,000 stocks, and the average return on the 50-stock sample is calculated.

- If this process were repeated several times with samples of the same size (50), the sample mean (estimate of the population mean) calculated will be different each time due to the different individual stocks making up each sample.
- The distribution of these sample means is called the **sampling distribution of the mean**.

Important

All the samples drawn from the population must be random, and of the same size.

The sampling distribution is different from the distribution of returns of each of the components of the population (each of the 10,000 stocks) and has different parameters.

Stratification is the process of grouping members of the population into relatively homogeneous subgroups, or strata, before drawing samples. The strata should be:

- *Mutually exclusive* i.e., each member of the population must be assigned to only one stratum.
- *Collectively exhaustive* i.e., no population element should be excluded from the sampling process.

Once this is accomplished, random sampling is applied within each stratum

- The number of observations drawn from each stratum is based on the size of the stratum relative to the population.

This often improves the *representativeness* of the sample by *reducing* sampling error.

Time-series data consists of observations measured over a period of time, spaced at uniform intervals.

- Example: The monthly returns on a particular stock over the last 5 years.

Cross-sectional data refers to data collected by observing many subjects at the same point in time. Analysis of cross-sectional data usually consists of comparing the differences among the subjects.

- Example: The returns of individual stocks over the last year.

Longitudinal data is data collected over time about multiple characteristics of the same observational unit.

- Example: The various economic indicators (multiple characteristics) of a particular country (observational unit) over a decade (period of time).

Panel data is data collected over time about a single characteristic of multiple observational units.

- Example: The unemployment rate (single characteristic) of a number of countries (multiple observational units) over time.

The **central limit theorem** allows us to make accurate statements about the population mean and variance using the sample mean and variance *regardless of the distribution of the population,* as long as the sample size is adequate.

Important Properties

- Given a population with *any* probability distribution, with mean, μ, and variance, σ^2, the sampling distribution of the sample mean, computed from sample size, *n*, will approximately be *normal* with mean, μ (the population mean), and variance, σ^2/n (population variance divided by sample size), when the sample size is greater than or equal to 30.
- No matter what the distribution of the population, for a sample whose size is greater than or equal to 30, the sample mean will be normally distributed).
- The mean of the population and the mean of the distribution of sample means are equal.
- The variance of the distribution of sample means equals σ^2/n, or population variance divided by sample size.

The standard deviation of the distribution of sample means is known as the **standard error** of the statistic.

When the Population Variance is Known

Example

A manufacturer claims that the life of the batteries that it produces is normally distributed with an average life of 30 hours, and a standard deviation of 5 hours. For a random sample of 40 batteries calculate the standard error.

When the Population Variance is Not Known

Example

A sample containing the monthly returns for 50 U.S. stocks has a mean of 5% and a standard deviation of 10%. Calculate the standard error of the sample mean.