Econometrics: Lecture 4

Hang Miao

Rutgers University

January 21, 2021

Overview

1. Population, Sample, Sampling Techniques
2. Estimation of unknown population parameters
3. Hypothesis testing
4. Confidence intervals

Population, Sample, Sampling Techniques

Population, Sample and Sampling

Definition (Population)

The collection of all elements possessing common characteristics that comprise universe is known as the population. It is the set of all elements under discussion for a given problem. Population represents the entirety of persons, units, objects and anything that is capable of being conceived.

Definition (Sample)

A subgroup of the members of population chosen for participation in the study is called sample.

Definition (Sampling)

Sampling is the selection of a sample from within a statistical population to estimate characteristics of the whole population. Statisticians attempt for the samples to represent the population in question. Two advantages of sampling are lower cost and faster data collection than measuring the entire population.

The Difference Between Population and Sample

- **Includes**: Population includes each and every unit of the universe under discussion. Sample only includes a proportion of the units in this universe.
- **Data collection**: For Population data is obtained by **Census** or a **Complete Enumeration** of all the values in the population. While sample are collected by **Sample Survey** or **Sampling**. o that one can make inferences or extrapolations from the sample to the population.
- **Characteristic**: The characteristic for population is called **parameters** while it is called **statistic** for sample.
- **Focus on**: the focus for population is to identify the characteristics of the elements; whereas in the case of the sample, the focus is made on making inferences or extrapolations from the sample to the population.

Types of Population and Sample

Different Types of Population

- **Finite Population**: The number of elements of the population is fixed and thus making it possible to enumerate it in totality.
- **Infinite Population**: The number of units in a population are uncountable, thus impossible to observe all the items of the universe.
- **Existent Population**: The population which comprises of objects that exist in reality.
- **Hypothetical Population**: Hypothetical or imaginary population is the population which exists hypothetically.

Different Types of Sample

- **Biased (Unrepresentative) Sample**: Some members of the intended population are less likely to be included than others.
- **Unbiased (Representative) Sample**: Each individual member of the population has a known, non-zero chance of being selected

Different Types of Sampling Techniques

- **Simple Random Sampling**: Each individual is chosen randomly and entirely by chance.
- **Systematic sampling**: Selection of elements from an ordered sampling frame. The sampling starts by selecting an element from the list at random and then every kth element in the frame is selected.
- **Stratified sampling**: Stratification is the process of dividing members of the population into homogeneous subgroups. Then simple random sampling or systematic sampling is applied within each stratum.
- **Cluster sampling**: The total population is divided into mutually homogeneous yet internally heterogeneous groups called clusters and a simple random sample of the groups is selected. If all elements in each sampled cluster are selected, then this is referred to as a "one-stage" cluster sampling. If a simple random subsample of elements is selected within each of these groups, this is referred to as a "two-stage" cluster sampling.

Estimation of unknown population parameters

Statistic, Estimator, Estimate

Definition (Statistic)

Statistic is a function of a sample data drawn from an unknown population. Statistic is random variable. The probability distribution of statistic is called **sampling distribution**. Statistic is not Estimator but Estimator is Statistic.

Definition (Estimator)

Estimator is a function of a sample data drawn from an unknown population related to some quantity of the underlying distribution of such population. Estimator is random variable.

Definition (Estimate)

Estimates are numeric values computed by estimators based on the sample data. Estimates are nonrandom numbers.

Example

Think of some economic variable, for example hourly earnings of college graduates, denoted by Y. Suppose we are interested in μ_Y the mean of Y. In order to exactly calculate μ_Y. we would have to interview every working graduate in the economy. We simply cannot do this due to time and cost constraints. However, we can draw a random sample of n i.i.d. observations $Y_1 \dots Y_n$ and estimate μ_Y using statistics like sample mean $\overline{Y}$. Then the estimator $\hat{\mu_Y}$ is defined as follow: .

$$\hat{\mu_Y} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

In this example, μ_Y is the parameter of the entire population characterize the central tendency of the hourly earning distribution. $\overline{Y}$ defined as $\frac{1}{n}\sum_{i=1}^{n} Y_i$ is the statistic. We use this $\overline{Y}$ statistic as the estimator $\hat{\mu_Y}$ to estimate the population parameter μ_Y

Evaluation for the Estimator

Definition (Unbiasedness)

An Estimator is unbiased if the expectation of the estimator equals to the population parameter which this estimator is intend to estimate. By our previous example, the estimator $\hat{\mu}_Y$ is unbiased if:

$$E(\hat{\mu}_Y) = \mu_Y$$

Definition (Consistency)

An estimator is consistent if as the sample size increases the estimates converge to the true parameter which the estimator is designed to estimate.

$$\hat{\mu}_Y \xrightarrow{p} \mu_Y$$

Definition (Efficiency)

For two unbiased estimator, the efficient estimator is the one has the lower variance.

Hypothesis testing

Hypothesis Testing

Definition (Hypothesis Testing)

Hypothesis test is a method of statistical inference. It is used for testing statistical relationship between the two data sets, the Sample data sets and the hypothetical Population.

Definition (Null Hypothesis)

The null hypothesis, denoted H_0 is the hypothesis we are interested in testing. In general, null hypothesis assumes no relationship between two measured phenomena, or no association among the sample and population.

Definition (Alternative Hypothesis)

The alternative hypothesis, denoted H_1 considered as the rival of the null hypothesis, it is the hypothesis that is thought to hold if the null hypothesis is rejected.

Hypothesis Testing

Definition (Type I error)

The Null Hypothesis is true but it is falsely rejected.

Definition (Type II error)

The Null Hypothesis is false but it is falsely assumed to be true.

Definition (Size of the Test)

The probability that the test rejects the true null hypothesis. i.e The probability the test made Type I error.

Definition (Power of the Test)

The probability that the test correctly rejects a false null. i.e The probability the test does not made Type II error.

Hypothesis Testing

Definition (Significant Level)

The significant level, α is a threshold probability under which the null hypothesis H_0 would be rejected in favor of the alternative H_1. It is the admissible risk of making a type 1 error, the statistical decision process can be controlled. In general, α would take on values like $10\%, 5\%, 1\%$

Definition (P-Value)

The P-Value of a result, p, is the probability of obtaining a result at least as extreme, given that the null hypothesis were true.

Definition (Statistically Significant)

The result is statistically significant at α level if $p \leq \alpha$.

Confidence intervals

Confidence Intervals

Definition (Confidence intervals)

Confidence Interval (CI) is a type of interval estimate, computed from the statistics of the observed data, that might contain the true value of an unknown population parameter.

Confidence Intervals for the Population Mean

$$99\% \text{ confidence interval for } \mu_Y = \left[\overline{Y} \pm 2.58 \times SE(\overline{Y})\right]$$
$$95\% \text{ confidence interval for } \mu_Y = \left[\overline{Y} \pm 1.96 \times SE(\overline{Y})\right]$$
$$90\% \text{ confidence interval for } \mu_Y = \left[\overline{Y} \pm 1.64 \times SE(\overline{Y})\right]$$