BECC-110

INTRODUCTORY ECONOMETRICS

School of Social Sciences
Indira Gandhi National Open University
Maidan Garhi, New Delhi-110068

cs and
ne

EXPERT COMMITTEE

COURSE PREPARATION TEAM

Block/ Unit Title		**Unit Writer**
Block 1	**Econometric Theory: Fundamentals**	
Unit 1	Introduction to Econometrics	Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi
Unit 2	Overview of Statistical Concepts	
Unit 3	Overview of Hypothesis Testing	
Block 2	**Regression Models: Two Variables Case**	
Unit 4	Simple Linear Regression Model: Estimation	Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi
Unit 5	Simple Linear Regression Model: Inferences	
Unit 6	Extension of Two Variable Regression Models	Prof. Kaustuva Barik, Indira Gandhi National Open University
Block 3	**Multiple Regression Models**	
Unit 7	Multiple Linear Regression Model: Estimation	Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi
Unit 8	Multiple Linear Regression Model: Inferences	
Unit 9	Extension of Regression Models: Dummy Variable Cases	Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi and Prof. B S Prakash, Indira Gandhi National Open University
Block 4	**Treatment of Violations of Assumptions**	
Unit 10	Multicollinearity	Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi
Unit 11	Heteroscedasticity	
Unit 12	Autocorrelation	
Block 5	**Econometric Model Specification and Diagnostic Testing**	
Unit 13	Model Selection Criteria	Dr. Sahba Fatima, Independent Researcher, Lucknow
Unit 14	Tests for Specification Errors	

Course Coordinator: Prof. Kaustuva Barik

Editors: Prof. Kaustuva Barik
Prof. B S Prakash (units 5, 7, 11-12) Saugato Sen (units 7-8)

PRINT PRODUCTION

September, 2021

ISBN:

Further information on the Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi, New Delhi -110068 or visit our website: http://www.ignou.ac.in

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi, by Registrar, MPDD.

Laser Typeset: Mr. Mukesh Yadav **Cover Design:** Mr. Sandeep Maini **Printed at:**

CONTENTS

COURSE INTRODUCTION

Econometrics is an interface between economics, mathematics and statistics. It is mainly concerned with the empirical estimation of economic theory. The present course provides a comprehensive introduction to basic econometric concepts and techniques. The course is divided into five blocks comprising 14 Units.

Block 1 titled, **Econometric Theory: Fundamentals**, comprises three units. Unit 1 is introductory in nature. It defines econometrics and lists the steps we follow in an econometric study. Unit 2 provides an overview of the concepts frequently used in econometrics. In Unit 3 we define the concept and procedure of hypothesis testing.

Block 2 is titled, **Regression Models: Two Variables Case**. It consists of three Units. Unit 4 begins with the estimation procedure of simple regression model by ordinary least squares (OLS) method. It also describes the properties of OLS estimators and goodness of fit of regression models. Unit 5 continues with the simple regression model and describes the procedure of testing of hypothesis. In this context it explains the procedure of forecasting with regression models. Unit 6 extends the simple regression models in terms of log-linear models and changing the measurement units of the variables in a regression model.

Block 3 titled, **Multiple Regression Models,** considers cases where there are more than one explanatory variable. There are three Units in this Block. Unit 7 deals with estimation of multiple regression models. Unit 8 deals with hypothesis testing in the case of multiple regression models. Unit 9 looks into structural stability of regression models and includes dummy variables as explanatory variables in multiple regression models.

Block 4 deals with **Treatment of Violations of Assumptions**. Unit 10 addresses the issue of multicollinearity. It outlines the consequences, detection and remedial measures of multicollinearity. Unit 11 deals with the issue of heteroscedasticity – its consequences, detection and remedial measures. Unit 12 deals with another important problem in multiple regression models, i.e., autocorrelation. It discusses the consequences, detection and remedial measures of autocorrelation.

Block 5 is titled, **Econometric Model Specification and Diagnostic Testing.** There are two Units in this Block. Unit 13 deals with model selection criteria. In this Unit we discuss issues such as the exclusion of relevant variables and inclusion of irrelevant variables. The subject matter of Unit 14 is tests for specification errors. In this context it gives an outline of Akaike Information Criterion (AIC), Schwarz Information Criterion (SIC), and Mallows’ Criterion.

UNIT 1 INTRODUCTION TO ECONOMETRICS [*]

Structure

1.0 OBJECTIVES

After going through this unit, you will be able to

- explain the significance of econometrics in the field of economics;
- distinguish between econometrics, mathematical economics and economic statistics;
- describe the steps to be followed in an econometric study; and
- distinguish between association and causation.

1.1 INTRODUCTION

Econometrics connects the real world to the existing economic theories. Econometrics is based on the development of statistical methods for testing economic relationships and various economic theories. Econometrics helps us in two ways so far as relationship among variables is concerned: (i) explaining the past relationship among the variables, and (ii) forecasting the value of one variable on the basis of other variables.

Econometrics is an interface between economics, mathematics and statistics. It is mainly concerned with the empirical estimation of economic theories. In a broad sense we can say that it is a branch of social science that combines the tools of mathematics and statistical inferences, and these tools are applied to analyse economic phenomena. Econometrics uses regression technique which establishes an association or relationship between various variables. You should note that such relationships do not imply causation. (i.e., cause and effect relationship). The notion of causation has to originate from some theory of economics.

[*] Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi

1.2 MEANING OF ECONOMETRICS

As mentioned earlier, econometrics deals with 'economic measurement'. It can be defined as a stream of social science which uses techniques of mathematics, statistical inference and economic theory applied to analyze any economic phenomenon. It deals with applications of mathematical statistics to economic data. The objective is to provide empirical support to the economic models constructed with the help of mathematical relationship and therefore obtain numerical results. Thus econometrics makes use of economic theory, mathematical economics, and economic statistics.

Econometrics hence becomes a platform for interaction of economic theory, (microeconomics or macroeconomics) using sophisticated mathematical tools in the form of mathematical equations and economic statistics, that is, data. Economic statistics is developed by collection, processing and presentation of data.

The central concern of mathematical economics is to express economic theory in mathematical forms or equations. These equations are finally are expressed in the form of models. You should note that mathematical economics does not evaluate the measurability or empirical verification of theory.

Economic statistics is primarily concerned with collection, processing and presentation of economic data in the form of charts, diagrams and tables. These data could be on microeconomic variables pertaining to households and firms or it could pertain to macroeconomic variables such as GDP, employment, prices, etc. Data for econometric models could be primary data or secondary data. An economic statistician usually limits himself/ herself to tabulation and processing of data.

Econometrics is mainly interested in empirical verification of economic theories. An econometrician would build models and test economic theories. In mathematical economics the relationship is deterministic. For example,

$$Y_i = a + bX_i \qquad \ldots(1.1)$$

In (1.1) above, Y is the explained variable

X is the explanatory variable

a and b are parameters.

The nature of relationship in econometrics, on the other hand, is stochastic. We add a stochastic error variable u_i to equation (1.1). For example,

$$Y_i = a + bX_i + u_i \qquad \ldots(1.2)$$

We will discuss further in Unit 4 on stochastic relationship among variables. In econometrics we generally require special methods due to the unique nature of economic data since such data are not generated under controlled experiments. The aim of econometrics is to bridge the gap between economic theory and actual measurement simply using the technique of statistical inference.

Thus, you should note three prominent features of econometrics. First, econometrics deals with quantitative analysis of economic relationships. Second, it is based on economic theory and logic. Third, it requires appropriate estimation methods to draw inferences. Thus, if the relationship is not expressed in quantitative terms we cannot apply econometric tools. Further, the variables are related according to some theory or logic; otherwise it will be similar to spurious correlation that you studied in statistics.

1.3 ECONOMICS AND ECONOMETIRCS

In economic theory the statements could be qualitative in nature. On the other hand, as discussed above, econometrics is a composition of mathematical economics, economic statistics and mathematical statistics. Let us take an example. The law of demand states that *ceteris paribus* (i.e., other things remaining the same) a rise in price of a commodity is expected to decrease the quantity demanded of that commodity. Therefore, economic theory predicts a negative or inverse relationship between price and quantity demanded of a commodity.

The law of demand does not provide any numerical measure of the strength of relationship between the two variables namely, price and quantity demanded of the commodity. It fails to answer the question that by how much the quantity will go up or down as a result of a certain change in price of commodity.

Econometrics provides empirical content to most economic theories. The real application of economics in the applied world includes forecasting various crucial economic variables such as sales, interest rates, money supply, price elasticity, etc.

The role of an economist is of great significance for an economy when it comes to understand how the variables would behave over a period of time or how these variables are connected to each other. An economist may be required to assess the impact of a proposed price increase on quantity demanded. For example, the impact of increase in price of electricity can be estimated by an econometrician and the electricity board may increase in price accordingly.

Check Your Progress 1

1) Bring out the differences between econometrics, mathematical economics and statistics.

....

....

....

....

2) Bring out the prominent features of econometrics.

....

....

....

....

1.4 METHODOLOGY OF ECONOMETRICS

In econometrics we generally come across several types of economic issues. These issues could be from any branch of economics such as microeconomics, macroeconomics, public economics, international trade, etc. These also could be from any of the sectors of the economy such as agriculture, industry and services. The problem at hand could be different. However, there certain common steps to be followed in an econometric study. These steps are as follows:

1. Construction of a statement of theory or hypothesis
2. Specification of mathematical model of the theory
3. Specification of statistical or econometric model
4. Obtaining requisite data
5. Estimation of the parameters of econometric model
6. Testing of hypothesis
7. Forecasting or prediction
8. Interpretation of results

These eight steps need to be elaborated further. Let us consider an example so that we can comprehend the issues. As you know from introductory macroeconomics, consumption expenditure depends upon income of households. Let us see how an econometric study can be carried out on the above relationship.

Step 1:Construction of a Statement of Theory or Hypothesis

The relationship between consumption and income is complex in nature. There are several factors that that influence consumption expenditure of a household such as size of family, education level, health status of family members, place of stay (rural/urban), etc. In a simple model, however, the Keynesian consumption function establishes the relationship between consumption expenditure and household income. There are two concepts used by Keynes: average propensity to consume (APC), and marginal propensity to consume (MPC). According to Keynes the APC has a tendency to decline as income level increases. We can take the above statement as a hypothesis. Recall that hypothesis is based on certain theory or logic.

Step 2: Specification of Mathematical Model of the Theory

The consumption function takes the following form:

$$C_i = C_0 + cY_i \qquad ...(1.3)$$

The variables C and Y represent consumption expenditure and income respectively. Note that C_0 is autonomous consumption, which is the bare minimum needed for survival. Even if income of a household is zero, consumption will be C_0. We note that for APC to decline, the parameters of equation (1.3) should fulfil the following two conditions: $C_0 > 0$ and $0 < c < 1$. These two conditions will help us in formulation of hypothesis in mathematical form.

Step 3: Specification of Statistical or Econometric Model

The consumption income relationship specified in equation (1.3) is exact in nature. If we plot the graph for equation (1.4) we will obtain a straight line. As mentioned earlier, the nature of relationship in econometrics is *stochastic*. Let us consider two households with the same level of income. Their consumption expenditure would be different due to certain factors other than income (such as health status of family members). In order to incorporate such factors we include another variable, u_i, in our model. The variable u_i has to meet certain conditions (to be discussed in Unit 4). Thus the econometric specification of the consumption function would be as follows:

$$C_i = C_0 + cY_i + u_i \qquad ...(1.4)$$

Step 4: Obtaining Requisite Data

Data can be obtained from primary sources or secondary sources. You should refer to Unit 1 of the course BECC 107: Statistical methods for Economics for details on primary data and secondary data. In that Unit we have discussed the procedure of conducting sample survey and the important sources of secondary data.

For estimation of our econometric model given at equation (1.4) we need data on two variables, viz., income (Y) and consumption expenditure (C). As you know, income and expenditure are flow variables. Thus we have to specify a time period for these variables. For convenience from measurement point of view, we can take monthly income and monthly expenditure. Second, we have to define what constitutes a household – who all are members of a household and who all are not included in the household. Third, we have to decide on the nature of data we collect.

As you know, four types of data are available. (i) time series, (ii) cross- sectional, (iii) pooled-data, and (iii) panel data.

(i) Time Series

Time series data are collected on a variable regularly over a period of time. There are some variables on which data is available on a daily basis (e.g., SENSEX and NIFTY). In the case of some other variables, it is available on monthly basis (e.g., consumer price index), on a quarterly basis (e.g., GDP) or on an annual basis (e.g., fiscal deficit).

(ii) Cross-Sectional Data

Cross-sectional data refers to data on several variables at a point of time. For example through a sample survey we can collect household data on expenditure, income, saving, debt, etc. Remember that time series data focuses on the same variable over a period of time while cross-sectional data focuses on several variables at the same point of time. Census data is an example of cross-sectional data.

(iii) Pooled Data

In the pooled data we have elements of both the time series and cross-sectional data. It is a time series of cross-sections. The observations in each cross section may not refer to the same unit. Let us consider an example. The census data in India is collected decennially. The number of households in each census however differs. Such data can be pooled to analyse the shifts in population characteristics over time. You can think of several other examples of pooled data. Examples could be employment and unemployment surveys, workforce participation rates, human development index, etc.

(iv) Panel Data

It is a special type of pooled data. Here observations are taken on the same sample units at multiple points of time. Suppose we want to analyse the variability of returns across shares in the stock market. We can take a sample of 50 public limited companies and observe their net asset value (NAV) daily for the month of August 2021. Thus we get 31 cross sections (since the month August has 31 days) of 50 firms. This constitutes a panel data. We call it a 'balanced panel' if all observations (for time period 1 to t; and for sample units 1 to n) are available. We call it an 'unbalanced panel' if some observations are missing.

Step 5: Estimation of the Parameters of the Econometric Model

We have discussed about sampling procedure, statistical estimation and testing of hypothesis in Block 4 of BECC 107. You need a thorough understanding of those concepts. Remember that in econometric estimation, the number of equations is more than the number of parameters. In order to estimate such models we need certain estimation methods. As you will come to know in subsequent Units of this course, there are quite a few estimation methods. You have been introduced to the least squares method in Unit 5 of the course BECC 107: Statistical Methods for Economics. There are certain econometric software available for estimation purpose. You will learn about econometric software in the course BECE 142: Applied Econometrics.

Step 6: Testing of Hypothesis

Once you obtain the estimates of the parameters, there is a need for test of the hypothesis. As you know, in a sampling distribution of an estimator, the estimate varies across sample. The estimate that you have obtained could be a matter of chance, and the parameter may be quite different from the estimate obtained. We need to confirm whether the difference between the parameter and the estimate really exists or it is a matter of sampling fluctuation.

For the consumption function (1.4), we should apply one sided t-test for testing of the condition $C_0 > 0$. For the marginal propensity to consume we should apply two-sided t-test $H_0 : c = 0$. For testing both the parameters together we should apply F-test.

There is a need to check for the correct specification of the model. Two issues are important here: (i) how many explanatory variables should be there in the regression model, and (ii) what is the functional form of the model.

The consumption function (see equation (1.4)) is a case of two-variable regress model. There is one explained variable and one explanatory variable in the model. If we include more number of explanatory variables (such as education, type of residential area, etc.) it becomes a multiple linear regression model. The functional form again could be linear or non-linear.

Step 7: Forecasting or Prediction

The estimated model can be used for forecasting or prediction. We have the actual value of the dependent variable. On the basis of the estimated regression model, we obtain the predicted value of the dependent variable. The discrepancy between the two is the prediction error. This prediction error is required to be as small as possible.

Step 8: Interpretation of Results

There is a need for correct interpretation of the estimates. In later Units of this course we will discuss issues such as model specification and interpretation of the result. The estimated model can be used for policy recommendation also.

1.5 ASSOCIATION AND CAUSATION

As you know from 'BECC 107: Statistical Methods for Economics' correlation implies association between two variables. Technically we can find out the correlation coefficient between any two variables (say the number of students visiting IGNOU library and the number of road accidents in Delhi). In some cases we find the correlation coefficients to be high also. Such relationship between variables however leads to spurious correlation. If we take two such variables (where correlation coefficient is high) and carry out a regression analysis we will find the estimates to be statistically significant. Such regression lines are meaningless. Thus regression analysis deals with the association or dependence of one variable on the other. It does not imply 'causation' however. The notion of causation has to come from existing theories in economics. Therefore a statistical relationship can only be statistically strong or suggestive. Unless causality is established between the variables the purpose of testing the economic theory would not make any sense. Most of the economic theories test the hypothesis whether one variable has a causal effect on the other.

Thus logic or economic theory is very important in regression analysis. We should not run a regression without establishing the logic for the relationship between the variables. Let us look into the case of the law of demand. While analysing consumer demand, we need to understand the effect of changing price of the good on the quantity demanded holding the other factors such as income, price of other goods, tastes and preferences of individuals unchanged. However, if the other factors are not held fixed, then it would be impossible to know the causal effect of price change on quantity demanded.

Check Your Progress 2

1) Explain the steps you would follow in an econometric study.

....

....

....

....

....

2) Assume that you have to carry out an econometric study on Keynesian consumption function. Write down the steps you would follow.

....

....

....

....

....

3) What do you understand by cause and effect relationship? How is it different from association?

....

....

....

....

....

1.6 LET US SUM UP

In this Unit we dealt with the significance of econometrics in the field of economics. Econometrics connects the real world with theory. It helps us to ascertain the validity of theory.

Behind every econometric model there should be certain logic. The relationship between variables should come from certain economic theory or logic. Mere estimation of a regression model may give up meaningless results.

In this Unit we described the steps of carrying out econometric analysis. There are eight steps that we should follow while conducting an econometric study.

1.7 ANSERS TO CHECK YOUR PORGRESS EXERCISES

Check Your Progress 1

1) In Section 1.2 we have shown that econometrics and interface between economics, statistics and mathematical economics. Elaborate on that.

2) There are three prominent features of econometrics. First, econometrics deals with quantitative analysis of economic relationships. Second, it is based on economic theory and logic. Third, it requires appropriate estimation methods to draw inferences.

Check Your Progress 2

1) You should explain the eight steps mentioned in Section 1.4.

2) You should follow the eight steps given in Section 1.4. Your answer may include the following:

   (i) Statement of the theory: $0 < MPC < 1$

   (ii) Mathematical specification of the model: $C = \beta_1 + \beta_2 Y$, $0 < \beta_2 < 1$

   (iii) Econometric specification the model: $C = \beta_1 + \beta_2 Y + u$

   (iv) Collection of Data: Secondary data from RBI Handbook of Statistics

   (v) Parameter Estimation: $\widehat{C_i} = -184.08 + 0.7164Y_i$

   (vi) Hypothesis Test: $\beta_1 > 0$ or $\beta_2 > 0$

   (vii) Prediction: what is the value of C, given the value of Y?

3) Regression analysis deals with the association or dependence of one variable on the other. It does not imply causation. The notion of causation has to come from outside statistics. It could be some existing theory in economics. Therefore a statistical relationship can only be statistically strong or suggestive. Most of the economic theories test the hypothesis whether one variable has a causal effect on the other. Regression *per se* is all about association between two or more variables; this association might be suggestive. Unless causality is established between the variables the purpose of testing the economic theory would not make any sense.

UNIT 2 OVERVIEW OF STATISTICAL CONCEPTS [*]

Structure

2.0 OBJECTIVES

After going through this unit, you will be able to

- explain the concept and significance of probability distribution;
- identify various types of probability distributions;
- describe the properties of various probability distributions such as normal, t, F and chi-square;
- explain the process of estimation of parameters and
- describe the properties of a good estimator.

[*] Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi

2.1 INTRODUCTION

Statistical concepts and estimation methods hold crucial significance in understanding the tools of econometrics. Therefore, you should be able to define various concepts and distinguish between them. The essence of econometrics is based on empirical analysis which deals with data. In fact, the tools of econometric analysis emerge from statistical methods.

Statistical concepts guide us to make judgement in the presence of uncertainty. Statistics provides the platform for data collection methods which becomes the basis for carrying out econometric analysis. Econometricians need to work with large population, which becomes a challenge. Therefore, there is a need to select appropriate sample and draw appropriate inferences based on probability distributions. Econometrics calls for a strong understanding of statistical concepts which help economists to choose the right sample and infer correctly from the chosen sample.

The population is a collection of items, events or people. It is difficult to examine every element in the population. Therefore it makes sense in taking a subset of the population and examining it. This subset of population is called a 'sample' which is further used to draw inferences. If the sample is random and large enough, the information collected from the sample can be used for making inference about the population.

Any experiment which gives random outcomes is referred to as random experiment. A variable which takes values which are outcome of random process is called a random variable. Thus, for a random variable each outcome is associated with certain probability of occurrence.

Random variables are discrete random variables when they take finite values. If the random variable assumes infinite number of values between any two pints, it is called a continuous random variable. Random variables have a probability distribution. If the random variable is discrete then the probability function associated with it is called 'probability distribution function'. If the random variable is continuous, then the probability function is referred to as 'probability density function'. Random variables can have variety of distribution functions depending on their probabilities. Some of the commonly used distribution functions are described in this Unit.

2.2 STATISTICAL INFERENCE

In BECC 107 we have discussed the procedure of statistical inference in detail (You should go through Units 13 and 14 of BECC 107). Statistical inference is the method of drawing conclusions about the population characteristics on the basis of information contained in a sample drawn from the population. Remember that population mean is not known to us, but we know the sample mean. In statistical inference we are interested in answering two types of questions. First, what would be the value of the population mean? The answer lies in making an informed guess about the population mean. This aspect of statistical inference is called 'estimation'. The second question pertains to certain assertion made about the population mean. Suppose a manufacturer of electric bulbs claims that the mean life of electric bulbs is equal to 2000 hours. On the basis of the sample information, can we say that the assertion is not correct? This aspect of statistical inference is called hypothesis testing. Thus statistical inference deals with two issues: (i) estimation, and (ii) hypothesis testing. We discuss about estimation of parameters in the present Unit. Hypothesis testing will be discussed in Unit 3.

If expected Price-Earning Ratio of 28 companies is 23.25, then this sample average can be used as an estimate of the population average of stocks. As you know, the sample average (or, sample mean) is denoted by $\bar{X}$. This sample mean can be inferred as the expected value of X, which is the population mean. This process of generalizing from the sample value ($\bar{X}$) to the population value E(X) is the essence of statistical inference.

Statistical inference aims at understanding the characteristics of population from the sample. These population characteristics are the 'parameters' of the population and the characteristics of the sample are the 'statistics'. The method of determining and computing population parameter using the sample is called *estimation*.

2.3 CENTRAL LIMIT THEOREM

When the functions of random variables are independent and identically distributed then as the sample size increases, the sample mean tends to be normally distributed around the population mean and the standard deviation reduces as sample size 'n' increases.

If X_1, X_2, X_3, ……. and X_n are independent and identically distributed with mean μ and standard deviation σ, then sample mean ($\bar{X}$) is given by

$$\bar{X} = \frac{(X_1 + X_2 + \ldots\ldots + X_n)}{n} \qquad \ldots(2.1)$$

The central limit theorem implies that the expected sample mean and standard deviation (SD) would converge as follows:

$$E\,(\bar{X}) = \mu \text{ and } SD\,(\bar{X}) = \frac{\sigma}{\sqrt{n}} \qquad \ldots\ (2.2)$$

The Central Limit Theorem (CLT) states that

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow N\,(0,\,1) \text{ as } n \rightarrow \infty \qquad \ldots\ (2.3)$$

This further implies that $\bar{X} \rightarrow$ N (μ, $\frac{\sigma^2}{n}$). In other words, the sample mean can be approximated with a normal random variable with mean μ and standard deviation $\frac{\sigma^2}{n}$. We discuss certain important probability distribution functions below.

2.4 NORMAL DISTRIBUTION

Normal distribution (also called z-distribution) is a continuous probability distribution function. This function is very useful because of Central Limit Theorem. It implies that averages of samples of observations of random variables independently drawn from independent distributions converge in distribution to the normal. It becomes normally distributed when the number of observations is sufficiently large. The normal distribution is also called the bell curve (see Fig. 2.1). The probability density function (pdf) of normal distribution is

$$f(\,x \mid \mu, \sigma^2\,) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{... (2.4)}$$

where, μ is the expectation of distribution or mean

σ is the standard deviation, and σ^2 is the variance.

Some of the important properties of normal distribution are:

a) The normal distribution curve is bell-shaped.

b) The normal curve is symmetrical about the mean μ.

c) The total area under the curve is equal to 1.

d) The area of the curve is completely described by its mean and standard deviation.

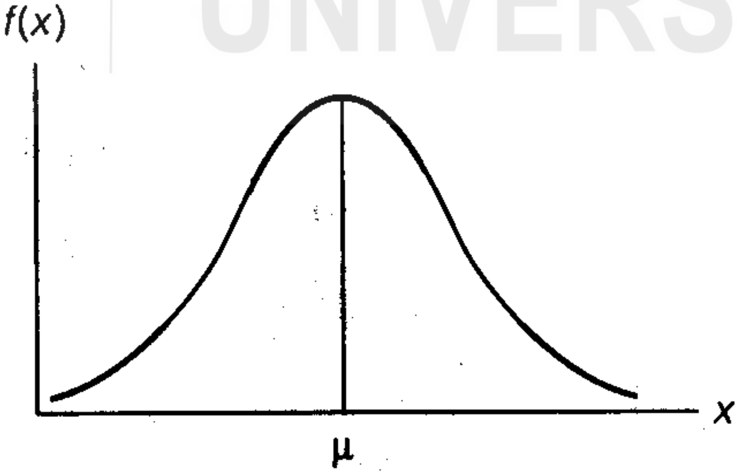

Fig. 2.1: Normal Probability Distribution

Standard Normal Distribution: N ~ N (0,1)

It is a normal distribution with mean zero ($\mu = 0$) and unit variance ($\sigma^2 = 1$), then the probability distribution function is given by

$$f(x \mid 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad \text{... (2.5)}$$

All the properties of the normal distribution mentioned above are applicable in the case of standard normal distribution.

Check Your Progress 1

1) Assume that X is normally distributed with mean $\mu = 30$ and standard deviation $\sigma = 4$. Find $P(X < 40)$.

....

....

....

....

....

....

2) Bring out the important properties of normal probability distribution.

....

....

....

....

....

....

3) The life of an electronic bulb produced by a company follows normal distribution with mean of 12 months and standard deviation of 2 months. Find out the probability that a bulb produced by the company will last

a) less than 7 months

b) between 7 and 12 months

....

....

....

....

....

....

2.5 CHI-SQUARE DISTRIBUTION

Suppose X is a normal variable with mean μ and standard deviation σ, then $z = \frac{X-\mu}{\sigma}$ is a standard normal variable, i.e., $z{\sim}\mathrm{N}(0,1)$. If we take the square of z, i.e., $z^2 = \left(\frac{\mathrm{X}-\mu}{\sigma}\right)^2$, then z^2 is said to be distributed as a χ^2 variable with one degree of freedom and expressed as χ_1^2.

It is clear that since χ_1^2 is a squared term; for *z* laying between $-\infty$ and $+\infty$, χ_1^2 will lie between 0 and $+\infty$ (because a squared term cannot take negative values). Again, since, *z* has a mean equal to zero, most of the values taken by *z* will be close to zero. As a result, the probability density of χ_1^2 variable will be maximum near zero.

Generalizing the result mentioned above, if $z_{1,}\, z_{2,\, \ldots,}\, .z_k$ are independent standard normal variables, then the variable

$$z = \sum_{i=1}^{ki} z_1^2$$

is said to be a χ^2 variable with k degrees of freedom and is denoted by χ_k^2. Fig. 2.2 given below shows the probability curves for χ^2 variables with different degree of freedom.

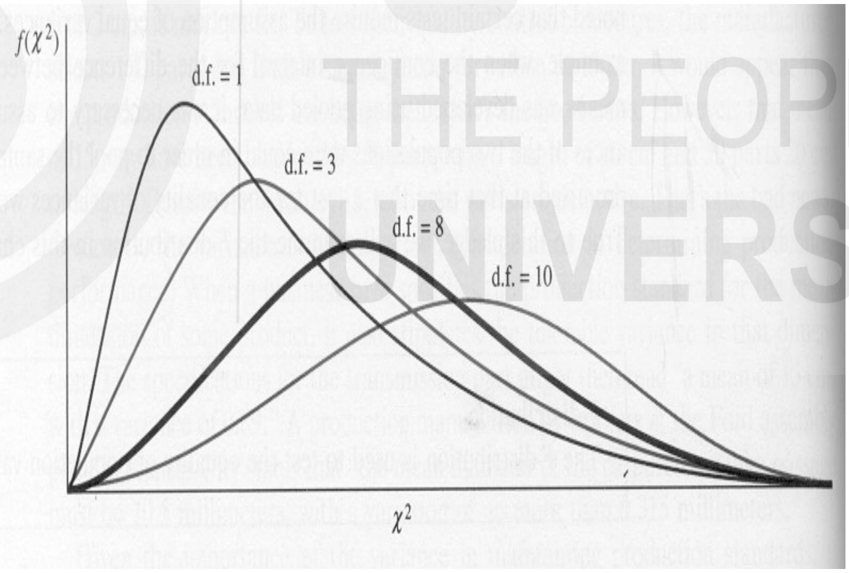

Fig. 2.2: Chi-Square Probability Curves

The chi-square distribution is one of the most widely used probability distributions. The area under the chi-square probability curve is equal to 1.

Unlike standard normal distribution, the distribution of chi-square changes its shape with sample size. In the case of small samples, the distribution is skewed to the right but it becomes symmetric as the sample size increases. All the values of the chi-square distribution are positive.

Properties of Chi-square distribution

1. The mean of the chi-square distribution is equal to the number of degrees of freedom (k).
2. The variance of the chi-square distribution is equal to two times the number of degrees of freedom: $\sigma^2 = 2k$
3. When the degrees of freedom are greater than or equal to 2, the maximum value of Y occurs when $\chi^2 = k - 2$.
4. As the degree of freedom increases, the chi-square curve approaches a normal distribution.

2.5 THE *t*- DISTRIBUTION

The t-distribution is also called the student's t-distribution. It was introduced by the English statistician W S Gosset under the penname 'Student'. It belongs to the family of continuous probability distributions. The t-distribution is applicable the sample size is small and population standard deviation is unknown. The cases where population parameters, i.e., μ and σ are not known and are estimated using sample statistics. The t-distribution is symmetric as in the case of the standard normal distribution (z). The height of the t-distribution depends on the sample size (see Fig. 2.3). As n approaches ∞, the t-distribution approaches the standard normal distribution.

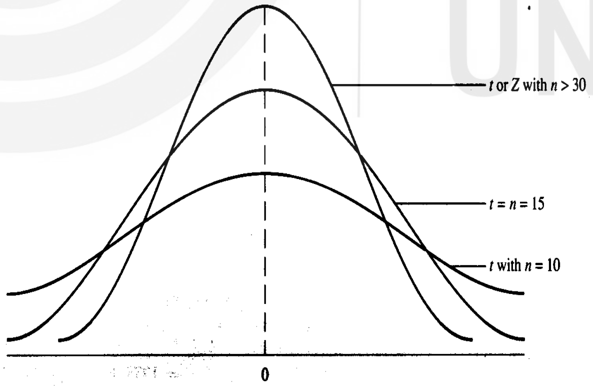

Fig 2.3: Student's-t Probability Curves

If z_1 is a standard normal variate, i.e., $z_1 \sim N(0,1)$ and z_2 is another independent variable that follows the chi-square distribution with *k* degrees of freedom, i.e.,

$z_2 \sim \chi_k^2$, then the variable

$$t = \frac{z_1}{\sqrt{(z_2/k)}} = \frac{z_1\sqrt{k}}{\sqrt{z_2}} \quad \dots (2.6)$$

is said to follow student's-*t* distribution with *k* degrees of freedom.

The value of t-distribution can be obtained as:

$$\text{t} = \frac{[\bar{X} - \mu]}{[\frac{s}{\sqrt{n}}]} \quad \dots (2.7)$$

where, $\bar{X}$ is the sample mean, μ is the population mean, s is the standard deviation of the sample and *n* is the sample size.

Properties of *t*-Distribution

1. The mean of the distribution is equal to 0
2. The variance is equal to [k/(k – 2)] where k is the degrees of freedom and k $\geq$2.
3. The variance is always greater than 1, although it is close to 1 when the degree of freedom is large. For infinite degrees of freedom the t-distribution is the same as the standard normal distribution.

The t-distribution can be used under the following conditions:

1. The population distribution is normal
2. The population distribution is symmetric, unimodal without outliers, and the sample size is at least 30
3. The population distribution is moderately skewed, unimodal without outliers and the sample size is at least 40
4. The sample size is greater than 40 without outliers.

Look into the above conditions. If the parent population (from which the sample is drawn) is normal we can apply t-distribution for any sample size. If population is not normal, the sample size should be large. The t-distribution should not be used with small samples drawn from a population that is not approximately normal.

2.7 THE *F*- DISTRIBUTION

Another continuous probability distribution that we discuss now is the *F* distribution. If z_1 and z_2 are two chi-squared variables that are independently distributed with $k_1$and k_2 degrees of freedom respectively, the variable

$$F = \frac{z_1/k_1}{z_2/k_2} \quad \dots (2.8)$$

follows *F* distribution with k_1 and k_2 degrees of freedom respectively. The variable is denoted by F_{k_1,k_2} where, the subscripts k_1 and k_2 are the degrees of freedom associated with the chi-squared variables.

You should note that k_1 is called the numerator degrees of the freedom and in the same way, k_2 is called the denominator degrees of freedom.

Some important properties of the F distribution are mentioned below.

1) The *F* distribution, like the chi-squared distribution, is also skewed to the right. But, as k_1 and k_2 increase, the *F* distribution approaches the normal distribution.

2) The mean of the *F* distribution is $k_1/(k_2 - 2)$, which is defined for $k_2 > 2$, and its variance is $\frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)}$ which is defined for $k_2 > 4$.

3) An *F* distribution with 1 and *k* as the numerator and denominator degrees of freedom respectively is the square of a student's-*t* distribution with *k* degrees of freedom. Symbolically,

$$F_{1,k} = t_k^2$$

4) For fairly large denominator degrees of freedom k_2, the product of the numerator degrees of freedom $k_1$and the *F* value is approximately equal to the chi-squared value with degrees freedom k_1, i.e., $k_1 F = \chi^2_{k_1}$.

The *F* distribution is extensively used in statistical inference and testing of hypotheses. Again, such uses also require obtaining areas under the *F* probability curve and consequently integrating the *F* density function. However, in this case also our task is facilitated by the provision of the *F* Table.

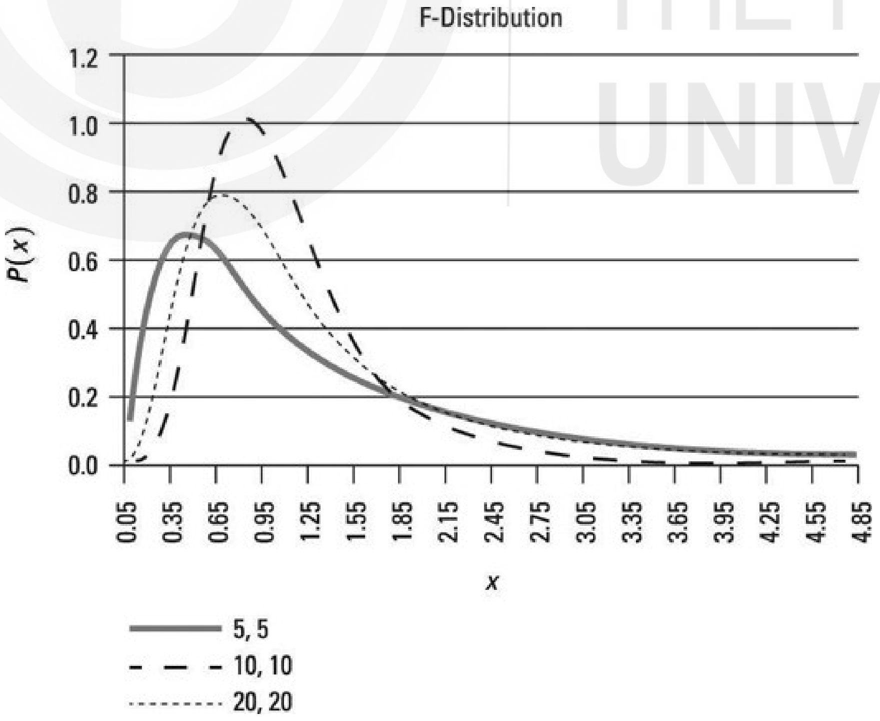

Fig. 2.4: Probability Curves of *F*-Distribution

The F-distribution is used to test the population variance. We can test whether two normal populations have the same variance. The null hypothesis is that the variances are same while alternative hypothesis is that one of the variances is larger than the other. That is:

$$\mathrm{H_{o:}}\ \sigma_1^2 = \sigma_2^2$$

$$\mathrm{H_{A:}}\ \sigma_1^2 > \sigma_2^2$$

The alternative hypothesis states that the first population has larger variance. The null hypothesis can be tested by drawing a sample from each population and calculating the estimates s_1^2 and s_2^2 . The samples are assumed to be independently drawn with size n_1 and n_2 respectively. We test the ratio

$$F = \frac{s_1^2}{s_2^2} \sim F_{n_{1-1,,}n_{2}\,_{-1}} \qquad \ldots (2.9)$$

If null hypothesis is not true, the ratio would be statistically different from unity. We should compare the calculated value of F (obtained from equation (2.9)) with the tabulated value of F (given in the appendix table at the end of the book). If the calculated value exceeds the tabulated value, then the null hypothesis is rejected.

Check Your Progress 2

1) A newly developed battery lasts 60 minutes on single charge. The standard deviation is 4 minutes. For the purpose of quality control test, the department randomly selects 7 batteries. The standard deviation of selected batteries is 6 minutes. What is the probability that the standard deviation in new test would be greater than 6 minutes?

....

....

....

....

....

....

2) Define chi-square distribution. Bring out its important properties.

....

....

....

....

....

....

3) Suppose the scores on a GRE test are normally distributed with population mean of 100. Suppose 20 people are randomly selected and tested. Sample standard deviation is 15. What is the probability that the average test score will be at most 110?

....

....

....

....

....

....

4) Test whether the students from private high schools are more homogeneous with respect to their science test score than the students from public high schools. It is given that the sample variances are 91.74 and 67.16 respectively for public and private schools. The sample sizes of the students are 506 for public schools and 94 for private schools.

....

....

....

....

....

2.8 ESTIMATION OF PARAMETERS

Estimation could be of two types: (i) point estimation, and (ii) interval estimation. In point estimation we estimate the value of the population parameter as a single point. On the other hand, in the case of interval estimation, we estimate the lower and upper bounds around the sample mean within which the population mean is likely to remain.

2.8.1 Point Estimation

Let us assume that a random variable X follows normal distribution. As you know, normal distribution is described by two parameters, viz., mean and standard deviation. Since we do not have data for the whole population (we have data for a sample only), we need to estimate mean E(X) = μ_X and variance σ_X^2 on the basis of a sample only.

Let us assume that we have data from a random sample of size *n* (suppose, sample size n = 50) from a known probability distribution (say, normal distribution). We use the sample to estimate the unknown parameters. Suppose, we find sample mean $\bar{X}$ to be 23.28. This single numerical value is called the point estimate of the parameter where $\bar{X} = \frac{\Sigma X_i}{n}$. This formula is called the point estimator. You should note that the point estimator is a random variable as its value varies from sample to sample.

2.8.2 Interval Estimation

In point estimation we estimate the parameter by a single value, usually the corresponding sample statistic. The point estimate may not be realistic in the sense that the parameter value may not exactly be equal to it.

An alternative procedure is to give an interval, which would hold the parameter with certain probability. Here we specify a lower limit and an upper limit within which the parameter value is likely to remain. Also we specify the probability of the parameter remaining in the interval. We call the *interval* as 'confidence interval' and the *probability* of the parameter remaining within this interval as 'confidence level' or 'confidence coefficient'.

The concept of confidence interval is somewhat complex. We have already explained it in BECC 107, Unit 13. Let us look at it again. We have drawn a sample of size *n* from a normal population. We do not know the population mean μ_X and population variance σ_X^2. We know the sample mean $\bar{X}$ and sample variance S_X^2. Since $\bar{X}$ varies across samples, we use the properties of the sampling distribution of $\bar{X}$ to draw inferences about μ_X.

If X is normally distributed, i.e. , we know that

$$\bar{X} \sim \left(\mu_X , \frac{\sigma_x^2}{n} \right) \qquad \ldots (2.10)$$

From (2.10) we can say that sampling distribution of sample mean $\bar{X}$ follows normal distribution with mean μ_X and standard deviation σ_X^2/n. Let us transform the above as a standard normal variable.

$$Z = \frac{\bar{X} - \mu_x}{\frac{\sigma_X}{\sqrt{n}}} \sim N(0, 1) \qquad \ldots (2.11)$$

Now the problem before us is that we do not know the population variance σ_X^2. Thus we take its estimator $S_X^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1}$. In that case, the appropriate test statistic is

$$t = \frac{(\bar{X} - \mu_X)}{S_x / \sqrt{n}} \qquad \ldots (2.11)$$

Equation (2.11) follows *t*-distribution with (*n* −1) degrees of freedom.

By re-arranging terms in equation (2.11) we obtain the confidence interval of μ_X.

This also helps us to obtain an interval estimation of μ_X.

For 27 degrees of freedom (d.f.), the tabulated value is 2.052 at the 5 per cent level of significance (see Appendix Table). Thus

$P(-2.052 \le t \le 2.052) = 0.95$ … (2.12)

The critical t values show the percentage of area under the t-distribution curve that remains between those values. The value $t = -2.052$ is called the lower critical value, and the value $t = 2.052$ is called the upper critical value.

Equation (2.12) implies that for 27 d.f. the probability is 0.95 or 95% that the interval (–2.052, 2.052) will include μ_X.

$$\therefore P\left(-2.052 \le t = \frac{\bar{X} - \mu_X}{S_x/\sqrt{n}} \le 2.052\right)$$

$$\Rightarrow \qquad P\left(\bar{X} - 2.052\frac{S_X}{\sqrt{n}} \le \mu_X \le \bar{X} + 2.052\frac{S_X}{\sqrt{n}}\right) = 0.95 \qquad (2.13)$$

Equation (2.13) provides an interval estimator of μ_X. It is called the 95% confidence interval (CI) for the true but unknown population mean μ_X. The value 0.95 is called the confidence coefficient. It implies that the probability is 0.95 that random interval $\bar{X} \pm 2.052\frac{S_X}{\sqrt{n}}$ contains true μ_X.

$\bar{X} - 2.052\frac{S_X}{\sqrt{n}}$ is called lower limit of interval.

$\bar{X} + 2.052\frac{S_X}{\sqrt{n}}$ is called upper limit of interval.

This is a random interval because the values are based on $\bar{X}$ and $\frac{S_X}{\sqrt{n}}$ which will vary from sample to sample. You should note that μ_X is not random; rather it is a fixed number. Therefore we can say that “the probability is 0.95 that μ_X lies in this interval”.

2.9 PROPERTIES OF ESTIMATORS

An estimator is considered as best linear unbiased estimator (BLUE) if it is linear, unbiased, efficient (with minimum variance). and also consistent implying that the the value of estimator converges to its true population value as the sample size increases. All the properties of good estimators are discussed below.

2.9.1 Linearity

An estimator is said to be a linear estimator if it is a linear function of the sample observation

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

$$= \frac{1}{n}\left(X_1 + X_2 + ... + X_n\right) \qquad \text{… (2.14)}$$

Sample mean is the linear estimator because it is a linear function of the observations.

2.9.2 Unbiasedness

The value of a statistic varies across samples due to sampling fluctuation. Although the individual values of a statistic may be different from the unknown population parameter, on an average, the value of a statistic should be equal to the population parameter. In other words, the sampling distribution of $\bar{X}$ should have a central tendency towards μ_X. This is known as the property of unbiasedness of an estimator. It means that although an individual value of a given estimator may be higher or lower than the unknown value of the population parameter, there is no bias on the part of the estimator to have values that are always greater or smaller than the unknown population parameter. If we accept that mean (here, expectation) is a proper measure for central tendency, then $\bar{X}$ is an *unbiased estimator* for μ_X if

$$E(\bar{X}) = \mu_X$$

2.9.3 Minimum Variance

An estimator of μ_X is said to be the minimum variance estimator if its variance is smaller than the variance of any other estimator of μ_X. Suppose there are three estimators of μ_X. The variance of $\hat{\mu}_3$ is the smallest of the three estimators. Hence, it is minimum variance estimator.

2.9.4 Efficiency

The property of unbiasedness is not adequate by itself. It is possible to obtain two or more estimators of a parameter as unbiased. Therefore, we must choose the most efficient estimator. Suppose two estimators of μ_X as given as follows:

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma^2}{n}\right) \qquad \text{… (2.15)}$$

$$X_{med} \sim N\left(\mu_X, \left(\frac{\pi}{2}\right)\frac{\sigma^2}{n}\right), \qquad \pi = 3.142 \text{ (approx.)} \qquad \text{… (2.16)}$$

In the case of large samples, the median computed from a random sample of normal population also follows normal distribution with the same μ_X. However, it has a large variance.

$$\frac{Var(\bar{X}_{med})}{Var(\bar{X})} = \frac{\pi}{2}\frac{\frac{\sigma^2}{n}}{\frac{\sigma^2}{n}} - \frac{\pi}{2} = 1.571 \qquad \text{(approx.)} \qquad \ldots (2.17)$$

Equation (2.17) implies that the variance of sample median is 57% larger than the variance of sample mean. Therefore, the sample mean provides more precise estimate of population mean compared to the median (X_{med}). Thus, $\bar{X}$ is an efficient estimator of μ_X.

2.9.5 Best Linear Unbiased Estimator (BLUE)

Suppose we consider a class of estimators. Among these estimators, an estimator fulfils three properties, viz., (i) it is linear, (ii) it is unbiased, and (iii) it has minimum variance. In that case, it is called a 'best linear unbiased estimator' (BLUE).

2.9.6 Consistency

Consistency is a large sample property. If we increase the sample size, the estimator should have a tendency to approach the value of the parameter. Thus, an estimator is said to be consistent if the estimator converges to the parameter as $n \to \infty$.

Suppose $X \sim N(\mu_X, \sigma_X^2)$. We draw a random sample of size n from the population.

Two estimators of μ_X are]

$$\bar{X} = \Sigma\frac{X_i}{n} \qquad \ldots (2.18)$$

$$X^* = \Sigma\frac{X_i}{n+1} \qquad \ldots (2.19)$$

As you know, the first estimator (2.18) is the sample mean and it is unbiased since $E(\bar{X}) = \mu_X$.

The second estimator (2.19) is biased as

$$E(X^*) = \left(\frac{n}{n+1}\right)\mu_X$$

Thus, $E(X^*) \neq \mu_X$

As the sample size increases we should not find much difference between the two estimators. As n increases, X^* will approach μ_X. Such an estimator is known as consistent estimator. An estimator is consistent estimator if it approaches the true value of parameter as sample size gets larger and larger.

Check Your Progress 3

1) Describe the desirable properties of an estimator.

....

....

....

....

2) For a sample of size 30, the sample mean and standard deviation are 15 and 10 respectively. Construct the confidence interval of population mean ($\boldsymbol{\mu_X}$) at 5 per cent level of significance.

....

....

....

....

2.10 LET US SUM UP

Statistical concepts guide us the way to make judgement in the presence of uncertainty. In this Unit we discussed about certain basic statistical concepts. We discussed about certain continuous probability distributions such as normal, standard normal, chi-square, *t* and *F*. We depicted the probability distribution curves of these curves. In the appendix given at the end of this book, we have given the following: Normal Area Table, and critical values of *t*, chi-square and *F* distributions.

In addition to the above we have described the properties of a good estimator. We have explained concepts such as unbiasedness, consistency and efficiency in the context of an estimator.

2.11 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) You have to find out the area under the standard normal curve.
If X = 40, $z = \frac{(40-30)}{4} = 2.5$
Hence P (X < 40) = P (z < 2.5) = [area to the left of 2.5] = 0.9938

2) Go through Section 2.4 and answer.

3) a) P(X < 7) = P (Z < – 2.5) = 0.0062

b) P(20 < X < 22) = P (–2.5 < Z < 0) = 0.4938

Check Your Progress 2

1) Standard deviation of the population is 4 minutes. Standard deviation of the sample is 6 minutes. The number of sample observations is 7.

   $X^2 = [(n–1)*s^2]/ \sigma^2$

   $X^2 = [(7 – 1) * 6^2] / 4^2 = 13.5$

   Degree of freedom is (n–1) = (7–1) = 6. The probability that a standard deviation would be less than or equal to 6 minutes is 0.96. This implies that the probability that the standard deviation would be greater than 6 minutes is (1 – 0.96) = 0.04.

2) Go through Section 2.5 and answer.

3) Population mean $\mu = 100$. Sample size n = 20. Degrees of freedom is (20–1) = 19. Sample mean $\bar{X}$ should be at most 110. Sample standard deviation s =15. Since we do not know the population standard deviation we apply t-distribution. Applying the formula,

   $$t = \frac{[\bar{X} - \mu]}{[\frac{s}{\sqrt{n}}]}\text{. Thus, } t = \frac{110 - 10}{\frac{15}{\sqrt{20}}} = 0.996$$

   This implies 99.6% chance that the sample average will be no greater than 110.

4) The degrees of freedom (n_1–1) and (n_2–1) are 505 and 93 respectively. Our null hypothesis H_0 is that the both type schools are equally homogeneous with respect to science marks. We are comparing variances. Thus we apply F-test.

   $$F = {s_1^2}/{s_2^2} = \frac{91.74}{67.16} = 1.366$$

   The tabulated value of F for 505 and 93 degrees of freedom is 1.27. Since calculated value is more than the tabulated value, we reject the H_0. We conclude that the students from private schools are more homogeneous with respect to science marks.

Check Your Progress 3

1) Go through Section 2.9 and answer.

2) Since population standard deviation is not known, you should apply t-distribution. Check the tabulated value of *t* given at the Appendix for 29 degrees of freedom and 5 per cent level of significance. Construct the confidence interval as given at equation (2.12).

UNIT 3 OVERVIEW OF HYPOTHESIS TESTING [*]

Structure

3.0 OBJECTIVES

After going through this unit, you will be able to

- explain the concept and significance of hypothesis testing;
- describe the applications of a test statistic;
- explain the procedure of testing of hypothesis of population parameters;
- distinguish between the Type I and Type II errors; and
- apply the tests for comparing parameters from two different samples.

3.1 INTRODUCTION

The purpose behind statistical inference is to use the sample to make judgement about the population parameters. The concept of hypothesis testing is crucial for predicting the value of population parameters using the sample. Various test statistics are used to test hypotheses related to population mean and variance. The variance of two different samples can also be compared using hypothesis testing. There are two approaches to testing of hypothesis: (i) test of significance

[*] Dr. Pooja Sharma, Assistant Professor, Daulat Ram College, University of Delhi

approach, and (ii) confidence interval approach. While testing a hypothesis, there is a likelihood of committing two types of errors: (i) type I error, and (ii) type II error. In this unit we will elaborate on the process of hypothesis testing, and explain the method of rejecting the null hypothesis on the basis of appropriate test statistic.

3.2 PROCEDURE OF HYPOTHESIS TESTING

We formulate a hypothesis on the basis of economic theory or logic. A hypothesis is a tentative statement about certain characteristic of a population. As you know, a population is described by its parameters (such as mean, standard deviation, etc.). Thus a hypothesis is an assumption about a population parameter. A hypothesis may or may not be true. For finding out that we test a hypothesis by certain econometric method.

Formulation of a hypothesis involves a prior judgement or expectation about what value a particular parameter may assume. For example, prior knowledge or an expert opinion tells us that the true average price to earnings (P/E) ratio in the local stock exchange is 20. Thus our hypothesis is that the P/E ratio is equal to 20.

In order to test this hypothesis, suppose we collect a random sample of stocks and find that the average P/E ratio is 23. Is the figure 23 statistically different from 20? Because of sampling variation there is likely to be a difference between a sample estimate and its population value. It is possible that statistically the number 23 may not be very different from the number 20. If this is the case, then we should not reject the hypothesis that the average P/E ratio is 20.

In hypothesis testing there are four important components: i) null hypothesis, ii) alternative hypothesis, iii) test statistic, and iv) interpretation of results. We elaborate on these components below.

(i) Formulation of null and alternative hypotheses: There are two types of hypothesis, viz., null hypothesis and alternative hypothesis. A 'null hypothesis' is the statement that we consider to be true about the population. It is called 'null' thereby meaning empty or void. For example, a null hypothesis could be: there is no relationship between employment and education. Therefore, if we carry out a regression of employment on education, the regression coefficient should be zero. Usually we denote null hypothesis by H_0. The alternative hypothesis is the opposite of the null hypothesis. Alternative hypothesis is usually denoted by H_1. You should note that H_0 and H_1 are 'mutually exclusive'; they cannot occur simultaneously.

(i) Identification of the test statistic: The null hypothesis is put to test by a test statistic. There are several test statistics (such as t, F, chi-square, etc.) available in econometrics. We have to identify the appropriate test statistic.

(ii) Interpretation of the results based on the value of the test statistic: After carrying out the test, we interpret the results. When we apply the test statistic to the sample data that we have, we obtain certain value of the test statistic (for example, t-ratio of 2.535). Interpretation of results involves comparison of two values: tabulated value of the test statistics and the computed value. If the computed value exceeds the tabulated value we reject the null hypothesis.

The sampling distribution of a test statistic under the null hypothesis is called the 'null distribution'. When the data depicts strong evidence against the null hypothesis, the value of test statistic becomes very large. By observing the computed value of the test statistic we draw inferences. Apart from the test statistic econometric software provides a *p-value*. The p-value indicates the probability of the null hypothesis being true. Thus, if we obtain a p-value of 0.04, it says the probability of the null hypothesis being true is 0.04 or 4 per cent. Therefore, if we take 5 per cent level of significance, we reject the null hypothesis.

3.3 ESTIMATION METHODS

In Unit 2 we described about two concepts; point estimation and interval estimation. We also discussed about certain probability distribution functions such as normal, *t*, *F* and chi-square.

There are basically three estimation methods: (i) least squares, (ii) maximum likelihood, and (iii) method of moments. We will use the least squares estimation method extensively in this course. In Unit 7 of this course we have introduced the maximum likelihood method. You are not introduced to 'Method of Moments' in this course.

In Unit 5 of the course BECC 107 we discussed with the concept of regression. In Section 5.9 that Unit we mentioned that the error variable in the regression should be minimised. For that purpose, we minimised the sum of squares of the error terms ($\sum u_i^2$). Now you can guess why it is called the least squares method. In this course we confine to ordinary least squares (OLS) method. We deal with OLS method first with the two-variable case. Subsequently, we extend it to more than two variables. This leads us the multiple regression model.

The name ordinary least squares (OLS) suggests that it is the simplest of the least squares methods. It implies that further complexities can be brought into the OLS method. Correctly so; there are generalised least squares (GLS), two-stage least squares (2SLS), three-stage least squares (3SLS), etc. Therefore, be careful when you read about the least squares method – notice which method the text is referring to. When you come across the term GLS in some context do not confuse it with OLS – both methods are different. In both OLS and GLS the sum of squares of the error terms is minimised (that is why both are referred to as least squares method) but there is some transformation of the regression model in the case of GLS. The advanced methods of least squares are not dealt with in this course. Remember that for carrying out the least squares method you do not need to assume any probability distribution function about the variables.

The maximum likelihood (ML) method assumes a probability distribution about the variables. Normal distribution is the most commonly used probability distribution function in maximum likelihood estimation. In ML method we form a likelihood function, which is derived from the probability distribution function. Note that in econometrics we are given the data – the data is obtained from a sample survey. We estimate the parameters of the regression model, under that the assumption that the data follows certain probability distribution function (for example, normal distribution). The likelihood function can follow any of the probability distribution functions; not just normal distribution. Recall from your statistics course that in probability distribution function we are given the parameters and we find out the probability of occurrence of particular dataset. In ML method, we do the opposite – we are provided with the data, and we are estimating the parameters.

The method of moments (MOM) makes use of the moment generating function (MGF) properties. You have been introduced to the concept of 'moments' in Unit 4 of BECC 107. The moment generating function of certain probability distributions are used for estimation of the parameters. The method of moments is quite advanced and beyond the scope of this course.

3.4 REJECTION REGION AND TYPES OF ERRORS

In the previous Unit we discussed about point estimation and interval estimation. The underlying idea behind hypothesis testing and interval estimation is the same. Recall that a confidence interval is built around sample mean with certain confidence level. A confidence level of 95 per cent implies that in 95 per cent cases the population mean would remain in the confidence interval estimated from the sample mean. It is implicit that in 5 per cent cases the population mean will not remain within the confidence interval. Note that when the population mean does not remain within the confidence interval our test statistic should reject the null hypothesis.

3.4.1 Rejection Region for Large Samples

Let us explain the concept of critical region. Sampling distribution of sample mean ($\bar{x}$) follows normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

The standard deviation of a sampling distribution is known as 'standard error'.

Thus, $\bar{x}$ can be transformed into a standard normal variable, z, so that it follows normal distribution with mean 0 and standard deviation 1.

In notations, $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ and $z \sim N(0,1)$.

Recall that area under the standard normal curve gives the probability for different range of values assumed by z. These probabilities are presented as the area under standard normal curve.

Let us explain the concept of critical region or rejection region through the standard normal curve given in Fig. 14.1 below. When we have a confidence coefficient of 95 percent, the area covered under the standard normal curve is 95 per cent. Thus 95 per cent area under the curve is bounded by $-1.96 \le z \le 1.96$. The remaining 5 per cent area is covered by $z \le -1.96$ and $z \ge 1.96$. Thus 2.5 per cent of area on both sides of the standard normal curve constitute the rejection region. This area is shown in Fig. 3.1. If the sample mean falls in the rejection region we reject the null hypothesis.

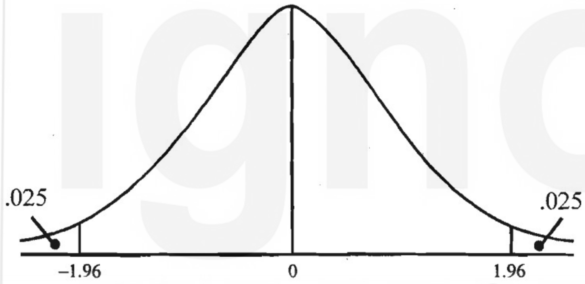

Fig. 3.1: Critical Regions

3.4.2 One-tail and Two-tail Tests

In Fig. 3.1 we have shown the rejection region on both sides of the standard normal curve. However, in many cases we may place the rejection region on one side (either left or right) of the standard normal curve. Remember that if α is the level of significance, then for a two-tail test $\frac{\alpha}{2}$ area is placed on both sides of the standard normal curve. But if it is a one-tail test, then α area is placed on one-side of the standard normal curve. Thus the critical value for one-tail and two tail test differ.

The selection of one-tail or two-tail test depends upon the formulation of the alternative hypothesis. When the alternative hypothesis is of the type $H_A : \bar{x} \ne \mu$ we have a two-tail test, because $\bar{x}$ could be either greater than or less than μ. On the other hand, if alternative hypothesis is of the type $H_A : \bar{x} < \mu$, then entire rejection is on the left hand side of the standard normal curve. Similarly, if the alternative hypothesis is of the type $H_A : \bar{x} > \mu$, then the entire rejection is on the right hand side of the standard normal curve.

The critical values for z depend upon the level of significance. In the appendix tables at the end of this book Table 14.1 these critical values for certain specified levels of significance (α) are given.

3.4.3 Rejection Region for Small Samples

In the case of small samples ($n \le 30$), if population standard deviation is known we apply z-statistic for hypothesis testing. On the other hand, if population standard deviation is not known we apply t-statistic. The same criteria apply to hypothesis testing also.

In the case of small samples if population standard deviation is known the test statistic is

$$z = \frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}} \qquad \ldots(3.1)$$

On the other hand, if population standard deviation is not known the test statistic is

$$t = \frac{|\bar{x} - \mu|}{s/\sqrt{n}} \qquad \ldots(3.2)$$

In the case of t-distribution, however, the area under the curve (which implies probability) changes according to degrees of freedom. Thus while finding the critical value of t we should take into account the degrees of freedom. You should remember two things while finding critical value of t. These are: i) level of significance, and ii) degrees of freedom.

3.5 TYPES OF ERRORS

In hypothesis testing we reject or do not reject a hypothesis with certain degree of confidence. As you know, a confidence coefficient of 0.95 implies that in 95 out of 100 samples the parameter remains within the acceptance region and in 5 per cent cases the parameter remains in the rejection region. Thus in 5 per cent cases the sample is drawn from the population but sample mean is too far away from the population mean. In such cases the sample belongs to the population but our test procedure rejects it. Obviously we commit an error such that H_0 is true but gets rejected. This is called 'Type I error'. Similarly there could be situations when the H_0 is not true, but on the basis of sample information we do not reject it. Such an error in decision making is termed 'Type II error' (see Table 3.1).

Note that Type I error specifies how much error we are in a position to tolerate. Type I error is equal to the level of significance, and is denoted by α. Remember that confidence coefficient is equal to $1-\alpha$.

The probability of committing a type I error is designated as α and is called the level of significance. The probability of committing type II error is called β. Thus, Type I error = α = prob (rejecting H_0 | H_0 is true)

Type II error = β = prob (accepting H_0 | H_0 is false)

Table 3.1: Type of Errors

	H_0 **true**	H_0 **not true**
Reject H_0	Type I Error	Correct decision
Do not reject H_0	Correct decision	Type II Error

Check Your Progress 1

1) Distinguish between one-tail and two-tail tests.

....

....

....

....

....

....

2) Distinguish between Type I and Type II errors.

....

....

....

....

....

....

3) Suppose the cholesterol level of an individual is normally distributed with mean of 180 and standard deviation of 20. Cholesterol level of over 225 is diagnosed as not healthy.

a) What is the probability of making type I error?

b) What level should people be diagnosed as not healthy if we want the probability of type I error to be 2%?

....

....

....

....

....

3.6 POWER OF TEST

As pointed out above, there are types I and type II errors in hypothesis testing. Thus, there are two types of risks: (i) α represents the probability that the null hypothesis is rejected when it is true and should not be rejected. (ii) β represents the probability that null hypothesis is not rejected when in reality it is false. The power of test is referred to as $(1 - \beta)$, that is the complement of β. It is basically the probability of not committing a type II error.

A 95% confidence coefficient means that we are prepared to accept at most 5% probability of committing type I error. We do not want to reject a true hypothesis by more than 5 out of 100 times. This is called 5% level of significance.

The power of test depends on the extent of difference between the actual population mean and the hypothesized mean. If the difference is large then the power of test will be much greater than if the difference is small. Therefore, selection of level of significance α is very crucial. Selecting large value of α makes it easier to reject the null hypothesis thereby increasing the power of the test $(1 - \beta)$.

At the same time increasing the sample size increases the precision in the estimates and increases the ability to detect the difference between the population parameter and sample, increasing the power of the test.

3.7 APPROACHES TO PARAMETER ESTIMATION

In statistical hypothesis testing, estimation theory deals with estimating the values of parameters based on measurement of empirical data that has a random component. The method of estimation requires setting up of a null hypothesis and a corresponding alternative hypothesis, which are further rejected or not rejected based on the two approaches used to make decision regarding the null hypothesis. The two methods have been described in the following section.

3.7.1 Test of Significance Approach

Any test statistic can be used for the test of significance approach to hypothesis testing. Let us consider the t-statistic.

$$t = \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \qquad \dots (3.3)$$

If the difference between $\bar{X}$ and μ_X is small, $|t|$ value will also be small, where $|t|$ is the absolute value of t-statistic. You should note that $t = 0$, if $\bar{X} = \mu_X$. In this case we do not reject the null hypothesis. As | t | gets larger, we would be more inclined to reject the null hypothesis.

Example. Suppose for a dataset $\bar{X} = 23.25, S_X = 4.49,$ and $n = 28$. Our nulland alternative hypothesis are

H_0: $\mu_X = 18.5$ and H_A: $\mu_X \neq 18.5$

$$\therefore \qquad t = \frac{23.25-18.5}{9.49/\sqrt{28}} = 2.6486 \qquad \ldots (3.4)$$

We need to specify α, the probability of rejecting the null hypothesis (probability of commuting type I error). Let us fix α at 5%.

$$H_0: \mu_X = 18.5$$

$$H_A:\ \mu_X \neq 18.5 \qquad \text{(two-tailed test)}$$

Since the computed *t* value is 2.6486. This value lies in the right-tail critical region of the *t*-distribution. We therefore reject the null hypothesis (H_0) that the true population mean is 18.5.

A test is statistically significant means that we one can reject the null hypothesis. This implies that the probability of observed difference between the sample value and the critical value (also called tabulated value) is not small and is not due to chance.

A test is statistically not significant means that we do not reject the null hypothesis. The difference between the sample value and the critical value could be due to sampling variation or due to chance mechanism.

3.7.2 Confidence Interval Approach

Let us assume that the level of significance or the probability of commuting type I error is fixed at α = 5%. Suppose the alternative hypothesis is two-sided. Assume that we apply *t*-distribution since variance is not known. From the *t* table we find the critical value of *t* at 8 degree of freedom $(n-K)=(10-2)$ at α = 5%. We find out the value to be 2.360. Thus we construct the confidence interval

$$P(-2.360 \leq t \leq 2.306) = 0.95 \qquad \ldots (3.5)$$

The probability that *t* value lies between the limits ($-2.360 \leq t \leq 2.360$) is 0.95 or 95%. The values -2.360 and 2.360 are the critical *t* values.

If we substitute the t from equation (3.2)

$$P\left(-2.306 \leq \frac{b_2-\beta_2}{\mathrm{SE(b_2)}} \leq 2.306\right) = 0.95 \qquad \ldots (3.6)$$

As we will see in Unit 4, $\mathrm{SE(b_2)}$ is $\frac{\hat{\sigma}}{\sqrt{\Sigma x_i^2}}$

If we substitute the above value in equation (3.6) and re-arrange terms we obtain

$$P\left(b_2 - 2.306\frac{\hat{\sigma}}{\sqrt{\Sigma x_i^2}} \leq \beta_2 \leq b_2 + 2.306\frac{\hat{\sigma}}{\sqrt{\Sigma x_i^2}}\right) = 0.95 \qquad \ldots (3.7)$$

Equation (3.7) provides a 95% confidence interval for the parameter $\beta_{2.}$ Such a confidence interval is known as the region of acceptance (H_0). The area outside the confidence interval is known as the rejection region (H_A).

If the confidence interval includes the value of the parameter β_2,we do not reject the hypothesis. But if the parameter lies outside the confidence interval, we reject the null hypothesis.

Check Your Progress 2

1) What is meant by power of a test?

....

....

....

....

....

....

2) Explain how a confidence interval is built.

....

....

....

....

....

....

3.8 LET US SUM UP

This unit elaborated the procedure of statistical inference regarding the population parameters. There are two approaches to hypothesis testing of population parameters: test of significance approach, and confidence interval approach. The unit also pointed out that there are errors involved in testing of hypothesis. While making a decision regarding acceptance or rejection of a hypothesis, two types of error may be committed: type I error, and type II error. Power of a test is the probability of not committing a type II error, i.e., rejecting H_0 when it is false is $(1 - \beta)$.

3.9 ANSWERS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Go through Sub-Section 3.4.2 and answer.

2) We have given the types of errors in table 3.1. You should elaborate on that.

3) a) In order to test this we use z-statistics $z = (X - \mu)/\sigma$, $z = (225 - 180)/20 = 2.25$

b) The area corresponding to the z value of 2.25 is 0.0122, which the probability of making type I error. An area of tail as 2% corresponds to $Z = 2.05$.

$Z = (X - \mu)/\sigma$

$2.05 = (X - \mu)/20$, i.e., $(X - \mu) = 2.05 * 20 = 41$

$X = 41 + 180 = 221$

Check Your Progress 2

1) Go through Section 3.6 and answer.

2) Go through Section 3.7.2 and answer.