PGEC-S2 03
Exam. Code: SM

Statistical Methods

SEMESTER - II

ECONOMICS

BLOCK 2

KRISHNA KANTA HANDIQUI STATE OPEN UNIVERSITY

Subject Experts

Professor Madhurjya Prasad Bezbaruah, Dept. of Economics, Gauhati University

Professor Nissar Ahmed Barua, Dept. of Economics, Gauhati University

Dr. Gautam Mazumdar, Dept. of Economics, Cotton University, Guwahati

Course Coordinator: Dr. Bhaskar Sarmah, Economics, KKHSOU

SLM Preparation Team

Units	Contributors
8 - 10	**Dr. Rijusmita Sarma**, HoD, Department of Statistics
	L. C. B. College, Guwahati
11	**Dr. Pranjal Sarma**, Department of Statistics
	L. C. B. College, Guwahati
12	**Bhaskar Sarmah**, KKHSOU
	Dr. Rupam Barman, Tezpur University
13, 14	**Dr. Mitali Deka,** NERIM, Guwahati

Editorial Team

Content	:	**Dr. Parag Dutta,** Associate Professor in Economics, KKHSOU (Units 8 - 12)
		Professor Tarakeswar Choudhary (Retired), Cotton College, Guwahati (Units 13-14)
Structure, Format & Graphics	:	**Dr. Bhaskar Sarmah,** Assistant Professor in Economics, KKHSOU
		Dr. Harekrishna Deka, Assistant Professor in Mathematics, KKHSOU

March, 2020.

ISBN: 978-93-89955-67-5

This Self Learning Material (SLM) of the Krishna KantaHandiqui State Open University is made available under a Creative Commons Attribution-NonCommercial-ShareAlike4.0 License (international): http://creativecommons.org/licenses/by-nc-sa/4.0/

Printed and published by Registrar on behalf of the Krishna Kanta Handiqui State Open University.

The university acknowledges with thanks the financial support provided by the **Distance Education Bureau, UGC, New Delhi**, for the preparation of this study material.

Headquarters: Patgaon, Rani Gate, Guwahati 781017; Web : www.kkhsou.in/web/

City Office: Housefed Complex, Dispur, Guwahati-781 006.

BLOCK INTRODUCTION

This is the second and last block of the course. This block consists of seven units, fron Unit 8 to 14. Unit 8 of the block gives us a broad overview of theoretical probability distributions relating to discrete variables. The next unit also deliberates on theoretical probability distributions. However this unit relates to continuous variables. Unit 10 discusses sampling and population. Various techniques of sampling have been discussed in this unit. The eleventh unit deliberates on discussing the characteristics of a good estimatro. After going through this unit, you will be able to discuss the most important characteristics of a good estimator. Unit 12 of the couse deliberates on index numbers. After going through this unit, you will be able to acquire knowledge about how index numbers are formed and the some of the most frequently index numbers. Unit 13 deals with testing of hypothesis. Here important concepts like null hypothesis and alternative hypothesis and the process of testing of hypotheis in case of large samples have been discussed. The last unit of the course, i.e., Unit 14 discusses about Test of Significance. Here, the concpets of testing of hypothesis in case case of small samples (less than 30 sample size) and the chi-square tests have been discussed.

On an overall, the study of the fourteen units of the course shall offer you a basic foundation of the course Statistics. We hope, this course will be of great help in your future and higher studies .

While going through a unit, you will notice some along-sode boxes, which have been included to help you know some of the difficult, unseen terms. Some ‘ACTIVITY’ (s) has been included to help you apply your own thoughts. Again, we have included some relevant concepts in ‘LET US KNOW’ along with the text. At the end of sections, you will get ‘CHECK YOUR PROGRESS’ questions. These have been designed to self-check your progress of study. It will be better if you solve the problems put in these boxes immediately after you go through the sections of the units and then match your answers with ‘ANSWERS TO CHECK YOUR PROGRESS’ given at the end of each unit.

CONTENTS

Pages :

UNIT 8: Theoretical Probability Distributions (Discrete Variable) **167-196**

Moment Generating Function: Definition, Generation of Moments, Properties of Moment Generating Function; Binomial Distribution: Derivation of Binomial Distribution, Definition, Moments of Binomial Distribution, Moment Generating Function of Binomial Distribution, Fitting of Binomial Distribution, Properties of Binomial Distribution; Poisson Process: Poisson Distribution as a Limiting Case of Binomial Distribution, Moments of Poisson Distribution, Moment Generating Function of Poisson Distribution, Fitting of Poisson Distribution, Properties of Poisson Distribution, Application of Poisson Distribution

UNIT 9: Theoretical Probability Distributions (Continuous Variable) **197-215**

Normal Distribution: Definition, Standard Normal Variate, Properties of Normal Distribution, Area under Normal Probability Curve, Importance of Normal Distribution; Weak Law of Large Numbers: Statement, Application of Weak Law of Large Numbers; Central Limit Theorem: Statement, Application of Central Limit Theorem

UNIT 10: Sampling and Population **216-242**

Concept of Population and Sample: Basic Principles of Sample Survey, Advantages of Sampling over Complete Enumeration, Different Steps in a Sample Survey, Errors in Surveys (Non-Sampling and Sampling Errors), Types of Sampling, Selection of Random Samples; Simple Random Sampling: Estimate of Population; Stratified Random Sampling: Advantage and Disadvantages of Stratified Random Sampling, Allocation in Stratified Random Sampling; Systematic Random Sampling: Linear Systimatic Sampling, Circular Systematic Sampling, Advantages and Disadvantages of Systematic Sampling, Uses of Systematic Sampling; Cluster Sampling: Advantages and Disadvanges of Cluster Sampling, Optimum Cluster size

Unit 11: Characterisitcs of a Good Estimator **243-255**

Estimators and Estimates; Types of Estimation; Criteria for a Good Estimator: Unbiasedness, Consistency, Efficiency

Unit 12: Index Numbers **256-278**

Index Numbers: Concept and Uses; Construction of Simple Index Numbers; Construction of Weighted Index Numbers: Laspeyres' Index Number, Paasche's Index Number, Fisher's Index Number; Test of Adequacy of Index Numbers: Time Reversal Test, Factor Reversal Test, Circular Test; Base Shifting, Splicing and Deflating; Index of Industrial Production (Concept only)

Pages :

Unit 13: Testing of Hypothesis **279-309**

Statistical Hypothesis: Null Hypothesis and Alternative Hypothesis; Errors in Hypothesis Testing, Level of Significance and Critical Region; One -Tailed and Two-Tailed Tests; General Procedure for Hypothesis Testing; Testing of Hypothesis in case of Large Samples: Hypothesis Testing for Single Population Mean, Hypothesis Testing for Single Population Proportion, Hypothesis Testing for Difference between Two Population Means, Hypothesis Testing for Difference between Two Population Means

Unit 14: Test of Significance **310-354**

The Chi-Square (χ^2) Distribution; Applications of χ^2 Test: χ^2 Test of Goodness of Fit, χ^2 Test for Independence of Attributes, χ^2 Test for a Specified Population Variance, χ^2 Test of Homogeneity; Testing of Hypothesis in case of Small Samples (t-Test): Characteristics of t-distribution, Applications of t-distribution, Hypothesis Testing for Single Population Mean, Hypothesis Testing for Difference between Two Population Means, Paired t-test for Difference between Two Population Means

References (For All Units of the Course) **355**

UNIT 8: THEORETICAL PROBABILITY DISTRIBUTIONS (DISCRETE VARIABLE)

UNIT STRUCTURE

8.1 Learning Objectives

8.2 Introduction

8.3 Moment Generating Function

- 8.3.1 Definition
- 8.3.2 Generation of Moments
- 8.3.3 Properties of Moment Generating Function
- 8.3.4 Illustrated Examples

8.4 Binomial Distribution

- 8.4.1 Derivation of Binomial Distribution
- 8.4.2 Definition
- 8.4.3 Moments of Binomial Distribution
- 8.4.4 Moment Generating Function of Binomial Distribution
- 8.4.5 Fitting of Binomial Distribution
- 8.4.6 Properties of Binomial Distribution
- 8.4.7 Illustrated Examples

8.5 Poisson Distribution

- 8.5.1 Poisson Distribution as a Limiting Case of Binomial Distribution
- 8.5.2 Moments of Poisson Distribution
- 8.5.3 Moment Generating Function of Poisson Distribution
- 8.5.4 Fitting of Poisson Distribution
- 8.5.5 Properties of Poisson Distribution
- 8.5.6 Application of Poisson Distribution
- 8.5.7 Illustrated Examples

8.6 Let Us Sum Up

8.7 Further Reading

8.8 Answers To Check Your Progress

8.9 Model Questions

8.1 LEARNING OBJECTIVES

After going through this unit, you will be able to-

- obtain moment generating function for any given distribution
- generate moments from any probability distribution
- learn some important properties of moment generating function
- derive binomial distribution from a coin tossing experiment
- derive poisson distribution as a limiting form of binomial distribution
- compute moments of binomial and poisson distribution
- fit binomial and poisson distribution to real life data
- describe the application of the concept of these probability distributions to real life data.

8.2 INTRODUCTION

So far we have brought out the nature of discrete and continuous probability distributions in general and also have defined their mean and variances. In this present unit we shall discuss about how the moments of a particular distribution as discussed in chapter 6, can be generated by a special kind of function called the moment generating function.

Again we know that the rule generating the probability distribution is the same whether we want to find the probability of the number of heads when three coins are lossed simultaueously or for the number of red balls in case of an urn containing 3 red and 2 black balls, out of which, say, two balls are drawn at random with replacement. Given this basic characteristic of a discrete distribution, our main concern is to examine the outcoms generated by an experiment and choose a distribution that best describes those outcomes.

Under this consideration we shall discuss some discrete probability distributions applicable to many real life situations in the subsequent sections of this unit. The distributions we are going to discuss here are binomial and poisson distributions.

8.3 MOMENT GENERATING FUNCTION

The moments of a probability distribution play an important role in theoretical as well as in applied statistics. With the help of moments we can measure the central tendency of a set of observations, their variability, their asymmetry and the height of the peak their curves would make. That is the knowledge of moments enable us to determine the probability distribution completely. So it would be very useful if a function could be found which will help in generating moments. One such function is moment generating function. So we can define a moment generating function as follows :

8.3.1 Definition

A function that is used to generate moments of a random variable X is called the moment generating function, denoted by Mx (t) and is given as

$$M_X(t) = E\left(e^{tx}\right).$$

When X is a discrete random variable with p.m. f. p(x)

$$M_x(t) = E(e^{tx}) = \sum_x e^{tx} p(x)$$

When x is a continous random variable in the interval (a,b) with p.d.f. f(x),

$$M_X(t) = E(e^{tx}) = \int_a^b e^{tx} f(x)dx \ .$$

8.3.2 Generation of Moments

We have, $M_X(t) = E\left(e^{tX}\right)$

$$= E(1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + ... + \frac{t^rX^r}{r!} + ...)$$

$$= E(1) + t\,E(X) + \frac{t^2}{2i}E(X^2) + \frac{t^3}{3i}E(X^3) + ... + \frac{t^r}{ri}E(X^r) + ...$$

$\therefore$ μ_1' = co-efficient of t in $M_X(t)$

μ_2' = co-efficient of $\frac{t^2}{2!}$ in $M_X(t)$

μ_3' = co-efficient of $\frac{t^3}{3!}$ in $M_X(t)$

Thus in general we see that,

$\therefore \mu_r'$ = co-efficient of $\frac{t^r}{r!}$ in $M_X(t)$.

Another convenient way of obtaining moments is to differentiate $M_X(t)$ successively and then put t = 0.

So if we differentiate $M_X(t)$, one time, with respect to t and then put t = 0, we get μ_1',

i.e. $\mu_1' = \left[\frac{d}{dt} M_X(t)\right]_{t=0}$

Differenting $M_x(t)$ 2 times with respect to t and then put t = 0 we get μ_α'. w

$\mu_2' = \left[\frac{d^2}{dt^2} M_X(t)\right]_{t=0}$

So in general, we get the raw moment of order r, μ_r' as

$\mu_r' = \left[\frac{d^r}{dt^r} M_X(t)\right]_{t=0}$

8.3.3 Preperties of Moment Generating Function

(i) If X is a random variable and c is a constant, there,

$M_{CX}(t) = M_X(ct)$

Proof : L. H. S = $M_{CX}(t) = E(e^{tcX})$

$= E(e^{ctX})$

$= M_X(ct)$

= R.H.S.

(ii) Additive property of moment generating function.

The moment generating function of sum of independent random variables is equal to the product of their moment generating functions. In symbol, if X_1, X_2, ..., X_n are independent random variables then,

$$M_{X_1+X_2+.....+X_n}(t) = M_{X_1}(t).M_{X_2}(t).....M_{X_n}(t)$$

Proof : L.H.S. = $M_{X_1+X_2+.....+X_n}(t)$

$= E\left[e^{t(x_1+x_2+...+x_n)}\right]$

$= E\left[e^{tx_1}.e^{tx_2}....e^{tx_n}\right]$

$= E(e^{tx_1}).E(e^{tx_2})...E(e^{tx_n})$ $\because X_1, X_2, ..., X_n$ are independent

= $M_{X_1}(t).M_{X_2}(t).....M_{X_n}(t)$

= R.H.S.

(iii) Effect of change of origion and scale on m. g. f.

$$M_{\frac{X-a}{h}}(t) = e^{\frac{-at}{h}} M_X\left(t/h\right)$$

Proof : $M_{\frac{X-a}{h}}(t) = E\left[e^{t\left(\frac{x-a}{h}\right)}\right]$

$= E\left[e^{\frac{tx}{h}} e^{\frac{-at}{h}}\right]$

$= e^{\frac{-at}{h}} E\left[e^{\frac{t}{h}(x)}\right]$

= $e^{\frac{-at}{h}} M_X\left(t/h\right)$

= R.H.S.

Note : (i) The moment generating function does not always exist.

(ii) In some cases the moment generating function exists, but it cannot generate moments.

8.3.4 Illustrated Examples

Example 1: Let X be a random variable with p. m. f.

$p(x) = q^x p$, x = 0, 1, 2, . . .

find the moment generating function and hance find its mean and variance.

Solution : $M_X(t) = E(e^{tx})$

$$= \sum_{x=0}^{\alpha} e^{tx} . q^x p$$

$$= p\sum_{x=0}^{\alpha} (qe^t)^x$$

$$= p\left[1 + qe^t + (qe^t)^2 + \ldots\right]$$

$$= p\left(1 - qe^t\right)^{-1} *$$

Mean = $\mu_1^{/} = \left[\frac{d}{dt} M_X(t)\right]_{t=0}$

$$= \left[(-1)p(1-qe^t)^{-2}(-qe^t)\right]_{t=0}$$

$$= pq(1-q)^{-2}$$

$$= \frac{pq}{p^2}$$

$$= \frac{q}{p}$$

$$\mu_2^{/} = \left[\frac{d^2}{dt^2} M_X(t)\right]_{t=0}$$

$$= \left[\frac{d}{dt}\left\{pq\ e^t(1-qe^t)^{-2}\right\}\right]_{t=0}$$

$$= pq\left[et(1-qe^t)^{-2} + e^t(-2)(1-qe^t)^{-3}(-qe^t)\right]_{t=0}$$

$$= pq\left[(1-q)^{-2} + 2q(1-q)^{-3}\right]$$

$$= \frac{pq}{p^2} + \frac{2pq^2}{p^3}$$

$$= \frac{q}{p} + \frac{2q^2}{p^2}$$

*The expansion used here is called binomial expansion with negative index, which is $(1-x)^{-1} = 1 + x + x^2 + x^3 + ...$

$\therefore$ variance $= \mu_2 = \mu_2^{/} - \mu_1^{/2}$

$$= \frac{q}{p} + \frac{2q^2}{p^2} - \frac{q^2}{p^2}$$

$$= \frac{q}{p} + \frac{q^2}{p^2}$$

$$= \frac{q(p+q)}{p^2}$$

$$= \frac{q}{p^2}$$

Example 2: Final moment generating function of a continuous random variable x having the p.d.f.

$$f(x) = \theta e^{-\theta x},\ x \geq 0$$

Hence find its mean and variance.

Solution : $M_X(t) = E(e^{tx})$

$$= \int_0^\infty e^{tx}\, \theta\, e^{-\theta x} dx$$

$$= \theta \int_0^\infty e^{(t-\theta)x} dx$$

$$= \theta \left[\frac{e^{(t-\theta)x}}{t-\theta} \right]_o^\infty$$

$$= \theta \left[\frac{1}{t-\theta}(0-1) \right]$$

$$= \frac{-\theta}{t-\theta}$$

$$= \frac{\theta}{\theta - t} = \left(1 - \frac{t}{\theta}\right)^{-1}$$

$$= 1 + \frac{t}{\theta} + \left(\frac{t}{\theta}\right)^2 + \left(\frac{t}{\theta}\right)^3 + \ldots.$$

$\therefore\ \mu_1'$ = mean - co-efficient of t. in $M_x(t) = \frac{1}{\theta}$

μ_2' = mean - co-efficient of t. in $M_x(t) = \frac{2}{\theta^2}$

$\therefore$ variance $= \mu_2 = \mu_2' - \mu_1'^2 = \frac{2}{\theta^2} - \left(\frac{1}{\theta}\right)^2 = \frac{1}{\theta^2}$



CHECK YOUR PROGRESS

Q 1: State true or false.

(i) Since M_x (t) generates moments, it is called moment generating function. (T / F)

(ii) If X an Y are independent random variables then

$$M_{X+Y}(t) = M_X(t) + M_Y(t) \qquad \text{(T / F)}$$

(iii) Moment generating function always exits. (T / F)

(iv) A moment generating function, if exists, determences the distribution function. (T / F).

8.4 BINOMIAL DISTRIBUTION

The binomial distribution, as derived by James Bernoulli in the year 1700, deals with population whose elements or outcomes can be devided into two catigories with reference to presence or absence of a particular attribute or characteristic. For example, the answer to a question may be correct or incorrect, results in an examination may be pass or fail, the factory workers may be educated or uneducated.

That is the outcome in each of these situations may be viewed in terms of presence or absence of an attribute. That is we consider only the two outcomes one possessing and the other not possissing the attribute. A convenient method is to denote them as ‘success’ and ‘failure’ according as the presence or absence of the attribute.

The mathematical background for binomial distribution describes descrete date resulting from an experiment called. Bernoulli trials, which are repeted finite number of times. Let us now device binomial distribution from a finite number of Bernoulli trials.

Note : A random experiment whoss outcomes has been classified into two categories, called ‘success’ and ‘failure’ is called Bernoulli trial. For example, tossing of a coin may result in a head or in a tail etc.

8.4.1 Derivation of Binomial Distribution

Consider a coin tossing experiment which is repeated, say, n number of times. When we get a head we call it a success and when we get a tail, we call it a failure, success is denoted by 's' and failure by 'f'. Probability of success is p and that of failure is q.

P(s) = p, P(f) = q, q = 1 – p.

suppose we are interested in getting x success out of n trials. Now the probability of x successes and consequently (n-x) failures in n independent trials, in a specified order, say, first x trials succes and the remaining (n-x) trials failures

P(ss . . . s f f . . . f) = P(s) P(s) . . . P(s) P(f) . . . P(f)

$\because$ the trials are independent

$$= \underbrace{p.p...p}_{x\ factor}.\ \underbrace{q.q...q}_{(n-x)\ factor}$$

$$= p^x q^{n-x}$$

But this is simply one way of getting the probabilety of x successes out of n trials. As we know, x successes in n trials can occur in n_{C_X} ways, and the probability for each of these cases is $p^x q^{n-x}$, Hence by the addition theorem of probability the probability of getting x success in n trials is given by.

$$p^x q^{n-x} + p^x q^{n-x} + ... + p^x q^{n-x},\ {}^n c_x \text{ times}$$

$$= {}^n c_x p^x q^{n-x}$$

Which is the probability mass function of binomial distribution. Again, x, the number of successes out of n trials can take values 0, 1, 2, ..., n. Therefore, we may define a binomial distribution as follows.

8.4.2 Definition

A random variable X is said to follow binomial distribution if it assumes non-zero values and if its p.m.f. is given by,

$$p(x) = p(X = x)\left\{n_{c_x}\ p^x\ q^{n-x},\ x = 0, 1, 2, ..., n\right\}$$

0, otherwise.

The two constante n an p are the parameters of this dist^{n}. It is also abbreviated as B(n, p).

Note : The probability of getting 0 success. 1 success, ..., n successes in this case, i.e. ${}^{n}c_0 p^0 q^n, {}^{n}c_1 pq^{n-1}, ..., {}^{n}c_n p^n q$ are the successive terms in the expansion of $(q+p)^n$, a binomial expansion, that is why this distribution is also termed as 'binomial distribution'.

8.4.3 Moments of Binomial Distribution

We are interested mainly in the first two moments of benomial distribution which are obtained as follows.

$$\mu_1' = Mean = E(X) = \sum_x x\, p(x)$$

$$= \sum_{x=0}^{n} x\ {}^{n}c_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x\ \frac{n}{x}{}^{n-1}c_{x-1} p^x q^{n-x} \qquad \because {}^{n}c_x = \frac{n}{x}{}^{n-1}c_{x-1}$$

$$= n\,p \sum_{x=1}^{n} {}^{n-1}c_{x-1} p^{x-1} q^{n-x}$$

$$= np\left[{}^{n-1}c_0\ p^0 q^{n-1} + {}^{n-1}c_1 pq^{n-2} + \ldots + {}^{n-1}c_{n-1} p^{n-1} q^0\right]$$

$$= np(q+p)^{n-1}$$

$$= np \qquad \because p+q=1$$

Hence the mean of binomial distribution is n p.

Again, $\mu_2' = E(X^2) = E\{X(X-1)\} + E(X)$

$$E\{X(X-1)\} = \sum_{x=0}^{n} x(x-1)\ {}^{n}c_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x(x-1)\frac{n(n-1)}{x(x-1)}\ {}^{n-2}c_{x-2} p^x q^{n-x}$$

$$= n(n-1)p^2 \sum_{x=2}^{n} {}^{n-2}c_{x-2} p^{x-2} q^{n-x}$$

$$= n(n-1)p^2 (q+p)^{n-2}$$

$$= n(n-1)p^2.$$

$\therefore E(x^2) = n(n-1)p^2 + np$

$\therefore \mu_2$ = variance $= \mu_2^{/} - \mu_1^{/2} = n(n-1)p^2 + n\,p - n^2p^2$

$= n^2p^2 - np^2 + np - n^2p^2$

$= np(1-p)$

$= npq$

So the variance of binomial distribution is npq .

Remark : For binomial distribution mean is always greater than variance. This is an important characteristic of binomial distribution.

8.4.4 Moment Generating Function of Binomial Distribution

If X ~ B (n, p), then

$$M_X(t) = E(e^{tx}) = \sum_x e^{tx} p(x)$$

$$= \sum_{x=0}^{n} e^{tx}\ {}^nc_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n} {}^nc_x (e^t p)^x q^{n-x}$$

$$= {}^nc_0 (e^t p)^0 q^n + {}^nc_1 (e^t p) q + \ldots + {}^nc_n (e^t p)^n q^0$$

$$= (q + e^t p)^n$$

which is the moment generating function of binomial dist^n.

8.4.5 Fitting of Binomial Distribution

The recurrence relation for probabilities of binomial distribution is, $p(x+1) = \frac{n-x}{x+1}\,\frac{p}{q}\,p(x)$. $\qquad x = 0,1,2,....,n-1$

Putting $x = 0,1,2,....,n-1$ we get $p(1), p(2),....., p(n)$. Here, we need to calculate only $p(0)$, which is.

$p(0) = {}^nc_0\ p^0 q^n = q^n$

For fitting a binomial distribution, we need to find the values of n and p . n is generally known, and $p = \frac{np}{n}$.

where np is the mean of the given set of data.

After calculating the value of p, we calculate the probabilities $p(1), p(2),....., p(n)$ as discussed above. Now multiplying these probabilities by N, the total frequency, we get the expected frequencies for x = 0, 1, 2, ..., n. Now we shall illustrate the fitting of biomial distribution with the help of an example.

Note : Calculation of expected frequency is called fitting.

Example 3: 3 coins are tossed 64 times and the following distribution of number of heads were obtained.

Number of heads	:	0	1	2	3
Frequency	:	10	18	20	16

Fit a binomial distribution under the hypothesis that the coins are unbiased.

Solution : Since the coins are unbiased

$p = q - \frac{1}{2}$ [in case coins are biased p is to be calculated from mean]

Here, N = 64, n = 3

Now, $p(0) = q^n = \left(\frac{1}{2}\right)^3 = \frac{1}{2^3} = \frac{1}{8}$

Now using recurrence relation, $p(x+1) = \frac{n-x}{x+1}\frac{p}{q}p(x)$

we calculate the probabilities as,

$$p(1) = \frac{3-0}{0+1}\frac{p}{q}p(0) = 3.\frac{1}{2^3} = \frac{3}{8}$$

$$p(2) = \frac{3-1}{1+1}\frac{p}{q}p(1) = \frac{3}{2}.\frac{3}{8} = \frac{3}{8}$$

$$p(3) = \frac{3-2}{2+1}\frac{p}{q}p(2) = \frac{1}{3}.\frac{3}{8} = \frac{1}{8}$$

Hence the expected frequenties are,

x	Expected frequency
0	$N \times p(0) = 64 \times \frac{1}{8} = 8$

1 $N \times p(1) = 64 \times \frac{3}{8} = 24$

2 $N \times p(2) = 64 \times \frac{3}{8} = 24$

3 $N \times p(3) = 64 \times \frac{1}{8} = 8$

8.4.6 Properties of Binomial Distribution

(i) The binomial distribution is a discrate distribution where the random variable X takes the values 0, 1, 2, ..., n.

(ii) The binomial distribution has two parameters, n and p .

(iii) The mean is equal to np and variance is given by npq .

(iv) The mean of binomial distribution is always greater than variance.

(v) If x_1 and x_2 follows binomial distribution with parameters (n_1, p) and (n_2, p) respectively, these x_1+ x_2 follows binomial distribution with parameter $(n_1 + n_2, p)$. THis is known as the additive property of binomial distribution.

(vi) Binomial distribution with parameters (n, p) tends to poisson distribution if n is very largs, p is very small and np is finite.

(vii) Binomial distributions also tends to Normal distribution if n is very large and neither nor q is very small.

8.4.7 Illustrated Examples

Example 4: A r. v. (r.v. : random variable) X follows binomial distribution with mean 5/3 and P(x = 2) = P (x = 1). Find variance, P(x = at least 1), P(x = at most 1).

Solution : If X follows binomial distribution with parameters n an p, the p.m.f. of X is,

$p(x) = {}^{n}c_{x} p^{x} q^{n-x}$, x = 0, 1, 2, . . . , n; $q = 1 - p$

Given, mean = $np = \frac{5}{3}$

and P(x = 2) = P(x = 1)

$$\Rightarrow {}^{n}c_2\, p^2 q^{n-2} = {}^{n}c_1\, p\, q^{n-1}$$

$$\Rightarrow \frac{n(n-1)}{2}\,\frac{p}{q} = n$$

$$\Rightarrow (n-1)p = 2q$$

$$\Rightarrow np - p = 2q$$

$$\Rightarrow \frac{5}{3} - p = 2q$$

$$\Rightarrow 5 - 3p = 6q$$

$$\Rightarrow 6q + 3p = 5$$

$$\Rightarrow 6(1-p) + 3p = 5$$

$$\Rightarrow 6 - 6p + 3p = 5$$

$$\Rightarrow -3p = -1$$

$$\Rightarrow p = 1/3$$

$$\therefore q = 1 - p = 2/3$$

Again, $np = \frac{5}{3}$ $\qquad\qquad \therefore\ n = \frac{5}{3} \Big/ \frac{1}{3} = 5$

Hence, variance $= npq = 5.\frac{1}{3}.\frac{2}{3} = \frac{10}{9}$

P(x = at least 1) = $p(x \geq 1) = 1 - p(x = 0) \because\ p(x = 0) + p(x \geq 1) = 1$

$$= 1 - {}^{5}C_0 \left(1/3\right)\left(2/3\right)^5 = 1 - \left(\frac{2}{3}\right)^5 = \frac{211}{243}$$

P(x = at most 1) = P(x=0) + p(x=0)

$$= {}^{5}C_0 \left(1/3\right)^0 \left(2/3\right)^5 + {}^{5}C_1 \left(1/3\right)^1 \left(2/3\right)^4$$

$$= \frac{112}{243}$$

Example 5: A box contains 3 red balls and 2 white balls. In a sample of 3 balls drawn, find the probability of getting (i) one red ball, (ii) two red balls, if the balls are drawn with replacement.

Solution : p = Prob of drawing a red ball

$$= \frac{3_{c_1}}{5_{c_1}} = \frac{3}{5}$$

$$\therefore q = 1 - p = \frac{2}{5}$$

$\therefore$ prob. of getting x red balls in a sample of 3 balls is,

$\therefore p(x) = 3_{C_x}\ p^x \quad p^{3-x}\ ,\ x = 0\ ,1\ ,2\ ,3$

Now, (i) P_x {drawing one red ball}

= p(1)

$= 3_{C_1}\left(\frac{3}{5}\right)^1\left(\frac{2}{5}\right)^2 = \frac{36}{125}$

(ii) p(2) $= 3_{C2}\left(\frac{3}{5}\right)^2\left(\frac{2}{5}\right)^1 = \frac{54}{125}$

Example 6: Ten coins are thrown simultaneously, Find the probability of getting at least seven heads.

Solution: p = prob. of getting a head = $\frac{1}{2}$

q = prob. of not gettig a head = $\frac{1}{2}$

$\therefore$ probability of getting x heads in a throw of 10 coins is

$\therefore p(x) = 10_{C_x}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{10-2} = 10_{C_x}\left(\frac{1}{2}\right)^{10}$, $x = 0,1,2,..,10$

$\therefore$ prob. of getting atleast 7 heads is given by,

$P(\geq 7) = p(7) + p(8) + p(9) + p(10)$

$= \left(\frac{1}{2}\right)^{10}\left\{10_{C_7} + 10_{C8} + 10_{C9} + 10_{C10}\right\}$

$= \frac{176}{1024}$

Example 7: The incidence of occupational disease in an industry is such that the workmen have a 20% chance of suffereing from it. What is the probability that out of six workmen, 4 or more will contract the disease?

Solution: prob that the workmen will suffer from

disease $= p = \frac{20}{100} = \frac{1}{5}$

$\therefore q = 1 - \frac{1}{5} = \frac{4}{5}$

n = 6

$\therefore$ using the probability law we get,

$$= P(x \geq 4)$$

$$= p(4) + p(5) + p(6)$$

$$= 6_{C_4}\left(\frac{1}{5}\right)^4\left(\frac{4}{5}\right)^2 + 6_{C_5}\left(\frac{1}{5}\right)^5\left(\frac{4}{5}\right) + 6_{C_6}\left(\frac{1}{5}\right)^6\left(\frac{4}{5}\right)^0$$

$$= \frac{1}{5^6}\left\{6_{C_4}.4^2 + 6_{C_5}.4 + 1\right\}$$

$$= \frac{1}{5^6}\left\{240 + 24 + 1\right\}$$

$$= \frac{265}{15625}$$

$$= 0.016.$$

Example 8: A multiple choice test consists of 10 questions with 3 answers to each questions, of which only one is correct. A student answers answers each question by rolling a dice and checking the first answer if he gets 1 or 2, second answer if the gets 3 or 4 and the third answer if he gets 5 or 6. To get a distinction, the student must secure at least 80% correct answer. If there is no negative marking, what is the probability that the student secures a distinction.

Solution: Probability of getting correct answer to a question is,

$$p = \frac{1}{3} \text{ and } q = \frac{2}{3}.$$

$\therefore$ Required probability of getting a distinction (i.e getting correct answers of 8 questions, out of 10) is,

$$P(x \geq 8) = P(x = 8) + P(x = q) + P(x = 10)$$

$$= 10_{C_8}\left(\frac{1}{3}\right)^8\left(\frac{2}{3}\right)^2 + 10_{C9}\left(\frac{1}{3}\right)^9\left(\frac{2}{5}\right) + 10_{C_{10}}\left(\frac{1}{3}\right)^{10}\left(\frac{2}{5}\right)^0$$

$$= \frac{1}{3^{10}}\left\{10_{C_8} \times 2^2 + 10_{C_9} \times 2 + 1\right\}$$

$$= \frac{1}{3^{10}}\left\{180 + 20 + 1\right\}$$

$$= \frac{201}{59049}$$

$$= 0.003$$

6. Probability of failure in statistics examination is 15%, 20 batchs of 5 students each appear for the examination. Obtain in how many batches 3 or more students would pass.

Sol^{n} : Given q = probability of failure = 15% = 0.15

$\therefore$ p = probability of success = 0.85

n = 5.

If X denote the number of students who would pass, then

$p(x)={}^{n}C_{x}p^{x}q^{n-x}$, x = 0, 1, . . ., 5.

We need $p(x \geq 3) = p(x=3) + p(x=4) + p(x=5)$

$= {}^{5}C_{3}p^{3}q^{5-3} + {}^{5}C_{4}p^{4}q^{5-4} + {}^{5}C_{5}p^{5}q^{0}$

= 10x $(.85)^3$ $(.15)^2$ + 5$(0.5)^4$ (0.15) + $(0.15)^5$

= 0.1381 + 0.3915 + 0.4437

= 0.9733.

$\therefore$ Required number of batches in which 30x more students would pass = 20 x 0.9733

= 19.468 $\approx$ 19.



CHECK YOUR PROGRESS

Q 2: If $X \sim B\left(12, \frac{1}{3}\right)$ what is E(X)?

Q 3: Examine the following statement and add your comment, "The mean of a binomial distribution is 5 and s.d. is 3".

Q 4: What are the parameters of binomial distribution?

Q 5: Find the varience of the binomial distribution

$p(x)={}^{10}C_{x}\left(\frac{2}{5}\right)^{x}\left(\frac{3}{5}\right)^{10-x}$, x = 0, 1, 2, . . ., 10.

Q 6: Define Bernoulli treats. Find its distributions and also its mean and variance.

Q 7: Derive Binomial distribution

or

Obtain the probability of exactly x successes in n independent Berioulli treats.

8.5 POISSON DISTRIBUTION

The Poisson process was first development by French mathematician and physicist serieon Devis Poisson, who published it in the year 1837. In the class of univariate distributions, the poisson distribution is perhaps the most useful one, because of its wids applicability in the description of natural phenomena, for example, demand for service in a big departmental store, major accident rate per day in a certain highway etc. A typical application of poisson distribution is for analysing queing (or waiting live) problems in which arriving customers during an interval of time arrive independently and the number of arrivals depends on the length of the time interval.

8.5.1 Poisson Distribution as a Limiting Case of Binomial Distribution

The binomial probability law of rare events gives rise to poisson probability law. That is poisson distribution can be derived as a limiting case of binomial distribution under the following conditions.

(i) n, the number of independent Bernoulli trials is very large, i, e $n \to \alpha$

(ii) p, the probability of success in each trial is very small, i.e. $p \to \alpha$

(iii) np = λ (say), is finite, where λ is a positive real number. under the above conditions, we can derive the p.m.f. of poisson distribution as follows:

Let us consider the p.m.f. of poisson distribution as,

$$p(x) = \binom{n}{x} p^x q^{n-x}, \qquad x = 0, 1, 2, \ldots, n; \qquad q = 1 - p.$$

Now, $$\lim_{n \to \alpha} p(x) = \lim_{n \to \alpha} \binom{n}{x} p^x q^{n-x}$$

$$= \lim_{n \to \alpha} \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \quad \because np = \lambda$$

$$= \frac{\lambda^x}{n!} \lim_{n \to \alpha} \frac{n!}{(n-x)!n^x} \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{\lambda^x}{n!} \lim_{n \to \alpha} \frac{n!}{(n-x)!n^x} \lim_{n \to \alpha} \left(1 - \frac{\lambda}{n}\right)^{n} \lim_{n \to \alpha} \left(1 - \frac{\lambda}{n}\right)^{-x}$$

$$= \frac{\lambda^x}{n!} \lim_{n \to \alpha} \frac{n(n-1)(n-2)\ldots(n-\overline{x-1})(n-x)!}{(n-x)!\, n^x} .1e^{-\lambda}$$

$$= \frac{\lambda^x}{n!} \lim_{n \to \alpha} \frac{n^x \left\{1.\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{x-1}{n}\right)\right.}{n^x}, e^{-\lambda}$$

$$= \frac{e^{-\lambda}\lambda^x}{x!} \qquad \because \lim_{n \to \alpha} \frac{1}{n} \to 0$$

Which is the p.m.f of poisson distribution. We give below the formal definition of poisson distribution as,

Definition :

A discrete random variable X is said to follow poisson distribution, if it assumes non -ve values only and its p.m.f is given by,

$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, x = 0, 1, 2, . . ., λ

= 0, λ is the only parameter of this distribution.

8.5.2 Moments of Poisson Distribution

The arithmatic mean of poisson distribution is given by,

$$\mu_1' = E(X) = \sum_x x\, p(x)$$

$$= \sum_{x=0}^{\alpha} x \frac{e^{-\lambda}\lambda^x}{x!}$$

$$= e^{-\lambda}\lambda \sum_{x=1}^{\alpha} \frac{x\, \lambda^{x-1}}{x(x-1)!}$$

$$= e^{-\lambda}.\lambda \sum_{x=1}^{\alpha} \frac{\lambda^{x-1}}{(x-1)!}$$

$$= e^{-\lambda}.\lambda\left[1+\lambda+\frac{\lambda^2}{2!}+\ldots\right]$$

$$= e^{-\lambda}.\lambda.e^{\lambda}$$

$$= \lambda .$$

$\therefore$ the mean of poisson distribution is λ.

$\mu_2^{/} = E(X^2) = E\{X(X-1)\} + E(X)$

Now, $= E\{X(X-1)\} = \sum_x x(x-1)p(x)$

$$= \sum_{x=0}^{\alpha} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!}$$

The exponential exponsion,

$$e^x = 1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\ldots$$

$$= e^{-\lambda}\lambda^2\sum_{x=2}^{\alpha} x(x-1)\frac{\lambda^{x-2}}{x.(x-1)(x-2)!}$$

$$= e^{-\lambda}\lambda^2\sum_{x=2}^{\alpha}\frac{\lambda^{x-2}}{(x-2)!}$$

$$= e^{-\lambda}\lambda^2\left[1+\lambda+\frac{\lambda^2}{2!}+\frac{\lambda^3}{3!}+\ldots\right]$$

$$= e^{-\lambda}\lambda^2.e^{\lambda}$$

$$= \lambda^2$$

$\therefore\ E(\lambda^2) = \lambda^2 + \lambda$

Now, $v(x) = \mu_2 = \mu_2^{/} - \mu_1^{/2}$

$$= \lambda^2 + \lambda - \lambda^2$$

$$= \lambda .$$

Hence the variance of poisson distribution is also λ.

Remark : Poisson distribution is the only discrate distribution where mean is equal to variance.

8.5.3 Moment Generating Function of Poisson Distribution

The moment generating function of poisson distribution is given by,

$$M_X(t) = E(e^{tx}) = \sum_{x=0}^{\alpha} e^{tx} . \frac{e^{-\lambda}\lambda^x}{x\,!}$$

$$= e^{-\lambda} \sum_{x=0}^{\alpha} \frac{\left(e^t \lambda\right)^2}{x\,!}$$

$$= e^{-\lambda} \left[1 + e^t \lambda + \frac{\left(e^t \lambda\right)^2}{2\,!} + \frac{\left(e^t \lambda\right)^3}{3\,!} + \ldots \right]$$

$$= e^{-\lambda}\, e^{e^t \lambda}$$

$$= -\lambda(1 - e^{+})$$

8.5.4 Fitting of Poisson Distribution

The p.m.f. of Poisson distribution with parameter is,

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x1},\ x = 0,1,2,.....$$

$$\therefore\ p(x+1) = \frac{e^{-\lambda}\lambda^{x+1}}{(x+1)!}$$

Now, $\frac{p(x+1)}{p(x)} = \frac{e^{-\lambda}\lambda^{x+1}/(x+1)!}{e^{-\lambda}\lambda^x / x!}$

$$= \frac{\lambda}{x+1} \qquad \because (x+1)!$$

$$= \ x\,.x!$$

Hence, the recurrence.....

The recurrence relation for probabilities of poisson distribution is,

$$p(x+1) = \frac{\lambda}{x+1} p(x), \qquad x \text{ = 0, 1, 2, ...} \qquad (1)$$

Putting x = 0, 1, 2, ... , in (1) we get the probabilities p(1), p(2), ... , etc. so we need to calculate p(0) only, which is,

$$p(0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-\lambda}\ .$$

where, λ is the mean of poisson distribution is to be obtained from data to be fitted.

Ultimately, the expected frequencies are calculated by

multiplying the corresponding probabilities by N, the total frequency.

We shall explain the fitting of poisson distribution to a given data set.

Example 9: 200 drivers made the following road accidents

Accidents :	0	1	2	3	4
No. of drivers : 122	60	15	2	1	

Fit a theoretical distribution to the data.

Solution : The suitable theoretical distribution to be fitted here may be the poission distribution,

Let, x_i = number of road accidents

f_i = no. of drivers.

Total number of drivers = N = 200

Here, λ = mean of poisson distribution

$$= \frac{1}{N} \sum_{i=1}^{n} f_i x_i$$

$$= \frac{100}{200} = 0.5$$

$\therefore\ p(0) = e^{-0.5} = 0.60653$

Using recurrence relation we get,

$$p(1) = \frac{\lambda}{0+1} p(0) = \lambda . p(0) = 0.5 \times 0.60653 = 0.30327$$

$$p(2) = \frac{\lambda}{2} p(1) = \frac{0.5}{2} \times 0.30327 = 0.07582$$

$$p(3) = \frac{\lambda}{3} p(2) = \frac{0.5}{3} \times 0.07582 = 0.01264$$

$$p(4) = \frac{\lambda}{4} p(3) = \frac{0.5}{4} \times 0.01264 = 0.00158$$

$\therefore$ the expected frequencies are,

$f(0) = N \times p(0)$ = 200 x 0.60653 = 121.306 $\approx$ 121

$f(1) = N \times p(1)$ = 200 x 0.30327 = 60.654 $\approx$ 61

$f(2) = N \times p(2)$ = 200 x 0.07582 = 15.164 $\approx$ 15

$f(3) = N \times p(3)$ = 200 x 0.01264 = 2.528 $\approx$ 3

$f(4) = N \times p(4)$ = 200 x 0.00158 = 0.316 $\approx$ 0

Total frequency = 200

8.5.5 Properties of Poisson Distribution

(i) Poisson distribution is a discrete distribution where the random variable X takes the values 0, 1, 2, ... , .

(ii) The poisson distribution has only one parameter λ.

(iii) The mean and the variance of poisson distribution are equal to λ.

(iv) The poisson distribution has only one or two modes.

(v) Poisson distribution can be obtained as a limiting case of binomial distribution.

(vi) If X and Y are independent poisson variates with paremeters P_1 and λ_2 respectively, then X +Y is also a poisson variate with parameter $\lambda_1 + \lambda_2$.

8.5.6 Application of Poisson Distribution

Physical situations in daily life where poisson distribution is applicable are,

(i) Number of telephone calls recieved per minute at a particular switch board.

(ii) Number of accidents in a city per unit time.

(iii) Number of defective parts produced in a factory per unit time.

(iv) Number of printing mistakes in each page of a book.

8.5.7 Illustrated Examples

Example 10 : Find the mean and variance of poisson distributionsuch that, P(x = 2) = P(x = 1)

Solution : It X follows poisson distribution with parameter λ,

then, $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, x = 0, 1, 2, ...

Now, p (x = 2) = p (x = 1)

$$\Rightarrow \frac{e^{-\lambda}\lambda^2}{2!} = \frac{e^{-\lambda}.\lambda}{1!}$$

$$\Rightarrow \frac{\lambda}{2} = 1$$

$$\Rightarrow \lambda = 2$$

$\therefore$ the mean and variance of poisson distribution is equal to $\lambda = 2$.

Example 11 : The number of ticketless travallers detected per checking follows poisson distribution with mean 10. Find the probability of getting two ticketless travellers.

Solution : Let X denote the no. of ticketless travellers in a checking. so X follows poisson distribution with given mean. $\lambda = 10$

$$\therefore p(x) = \frac{e^{-10}\, 10^x}{x\,!}, \qquad x = 0, 1, 2, \ldots .$$

$\therefore$ P(getting two ticketless travellers during a checking)

= P(x = 2)

$$= \frac{e^{-10}\, 10^2}{2\,!}$$

= 0.0025.

Example 12: What probability model is appropriats to describs a situation where 100 misprints are distributed randomly throughout the 100 pages of a book? For this model what is the probability that a page observed at random will contain at least 3 misprints.

Solution : Since 100 misprints are distributed randomly throughout the 100 pages of the book, the probability that there being a misprint is $p = \frac{1}{100} = 0.01$, which is very small.

n = 100 pages is quite large.

Hence poisson distribution is best fitted in this case.

Here, $\lambda = np = 0.01 \times 100 = 1$

$\therefore$ Probability that there are alteast three misprints in a page is

$$P(x \geq 3) = 1 - p(x < 3) = 1 - \{P(x = 0) + P(x = 1) + P(x = 2)\}$$

$$= 1 - \left\{ e^{-1} + e^{-1} + \frac{e^{-1}}{2!} \right\}$$

$$= 1 - 2.5 e^{-1}$$

$$= 1 - 2.5 \times 0.3679$$

$$= 0.0802 \cdot$$

Example 13: A factory producting machine parts finds that on an average there are 5 defective items in a lot of 1000 units. The units are packed, 500 in packet. Calculate the number of packets containing 3 defective and 4 defectives in a consignment of 2000.

Solution: Here n = 500, is large

p = $\frac{5}{1000} = .005$, is very small.

N = 2,000.

Hence, the distribution can be taken as Poisson distribution.

Here, $\lambda = np = 500 \times \frac{5}{1000} = 2.5$

Again, the number of packets with x defectives is

given by, $N.\frac{e^{-\lambda}\lambda^x}{x!}$

$\therefore$ No. of packets with 3 defectives

$$= \frac{N.e^{-\lambda}\lambda^3}{3!} = 2000 \times \frac{e^{-2.5}(2.5)^3}{3!}$$

=

No. of packets with 4 defectives

$$= \frac{N.e^{-\lambda}\lambda^4}{4!} = 2000 \times \frac{e^{-2.5}(2.5)^4}{4!}$$

Example 14: A car fire firm has two cars which it hired out day by day. The number of demands for a car on each day is distributed as Poisson variate with mean 1.5. Calculate how many days in a month (i) neither car is used and (ii) some demand is refused.

Solution: Probability of days on which there are x demands for a car = p(x)

$$= \frac{e^{-1.5}(1.5)^x}{x!}$$

since the number of demands for a car on any day is a Poisson variate with mean 1.5.

(i) the number of days in a month, on which neither car is used is given by,

$$30 \times p(0)$$

$$= 30 \times \frac{e^{-1.5}(1.5)^0}{01}$$

$= 30 \times .2231$

$= 6.69$

$\cong 7$

(ii) No. of days in a month, when some demand is refused is given by, $30 \times P(x > 2)$

$$= 30 \times \{1 - P(x \le 2)\}$$

$$= 30 \times [1 - \{P(x = 0) + P(x = 1) + P(x = 2)\}]$$

$$= 30 \times \left[1 - e^{1.5}\left\{1 + 1.5 + \frac{(1.5)^2}{2!}\right\}\right]$$

$$= 30 \times \{1 - (0.2231 \times 3.625)\}$$

$$= 30 \times 0.19126$$

$$= 5.7 \approx 6$$



CHECK YOUR PROGRESS

Q 8: Fill in the blanks :

(i) For a poisson distribution mean is ..... to variance (less/equal/greater)

(ii) The s.d. of a poisson distribution with mean 4 is ..... (4/2).

(iii) For a poisson distribution $\frac{e^{-1.5}(1.5)^x}{x!}$, x = 0, 1, 2, ... the variance is (1.5/2.5).

(iv) If x follows poisson distribution then its m.g.f is $\left(e^{-\lambda(1-e^t)} / e^{-\lambda}(1-e^t)\right)$

(v) If $x_i = 1, 2, ..., n$ are n poisson variates then the distribution of $\sum_{i=1}^{n} x_i$ is also a ..... variate. (Binomial / Poisson)

Q 9: "Poisson distribution is also known as probability distribution of rare events" – Explain.

Q 10: Give some examples of poisson distribution.



8.6 LET US SUM UP

- A function, which is used to generate moments of a probability distribution is known as moment generating function.
- The importance of moment generating function lies in the fact that, a probability distribution can be completely determined if its moments are known.
- The moment generating function of a random variable X is given by, $M_X(t) = E(e^{tx})$, t > 0.
- A random experiment whose outcomes are classified into two categories, called 'success' and 'failure' is called a Bernoulli trials.
- The p. m. f. of binomial distribution with parameters n and p is, $p(x) = {}^nC_x p^x q^{n-x}$, x = 0, 1, 2, . . ., n.
- The mean of binomial distribution is np and varience is npq.
- For a binomial distribution, mean is always greater than variance.
- Sum of independent binomial variates is also a binomial variate.
- Poisson distribution can be derived as a limiting case of binomial distribution when, $n \to \alpha$, $p \to o$ and $np \to \lambda$ (say).
- The p. m. f. of poisson distribution with parameter λ is given by, $p(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, x = 0, 1, 2, ... , λ.
- Poisson distribution plays an important role in the discription of random phenomena in a continuous interval of any kind. For example, telephone calls at a telephone exchange within a specified period of time.
- Sum of independent poisson variates is also a poisson variate.



8.7 FURTHER READING

1) Das, K. K, Bhattacharjee, D (2008). A Treatise on statistical Inference and Distributions. Asian Books Pvt. Ltd.

2) Gupta, S. C. and Kapoor, V. K. (2002). Fundementals of Mathematical statistics; Sultan Chand & Sons.

3) Goon, A. M. Gupta. M. P and Dasgupta, B (2002). Fundementals of statistics (Vol I), Seventh Revised Edition; The World Press Pvt. Ltd. Kolkata.

4) Saxena, H. C. (1994). Probability and its application, sixth edition; S. Chand and Company, New Delhi.

5) Singh, R (2012). An Introduction to Probability and Probability Distributions, Books and Ellied (P) Ltd.



8.8 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: (i) True, (ii) False, (iii) False, (iv) True.

Ans to Q No 2: We know that if X ~ B (n, p) then E(x) = np.

Here, n = 12, $p = \frac{1}{3}$

$\therefore\ E(x) = np = 12.\frac{1}{3} = 4$

Ans to Q No 3: Here, given, mean = 5, s. d. = 3

$\therefore$ variance = 9

But we know that for a binomial distribution, mean is always greater than variance. Hence, the above statement is false.

Ans to Q No 4: The parameters of a binomial distribution are n and p.

Ans to Q No 5: Given, $p(x) = {}^{10}C_x \left(\frac{2}{5}\right)^x \left(\frac{3}{5}\right)^{10-x}$, x = 0, 1, 2, ... , 10.

Here, n = 10, $p = \frac{2}{5}, \quad q = \frac{3}{5}$

Hence, variance = $npq = 10.\frac{2}{5}.\frac{3}{5} = \frac{12}{5}$

Ans to Q No 6: A particular trial having two outcomes only, viz success and failure occurring with probability p and (1–p) respectively is called a Bernoulli trial. e.g. tossing of a coin.

A random variable X is said to follow Beruoulli distribution if its p.m.f. is given by,

$p(x) = p^x(1-p)^{1-x}, \quad x = 0, 1$.

p is the only parameter of this distribution

$\therefore$ Mean = E(x) = 0. (1–p) + 1 . p = p.

$V(x) = E(x^2) - \{E(x)^2\}$.

$E(x^2) = 0^2 . (x-p) + 1^2 . p = p$

$\therefore V(x) = p - p^2 = p(1-p)$.

Ans to Q No 7: Consider n independent Bernoulli trials each having two and two outcomes only called success and failure, Probability of success is p and the probability of failure is (1 – p) = q.

The probability of x successes and (n–x) failures in n independent trials in a specific order, say, s s s... f.f.f is given by,

P (s. s. . . . s. f . f. . . . f) = p(s) p(s) s . . . p(s) p(f) . . . p(f)

$\because$ trials are independent

= p . p . . . p q q . . . q

x times (n-x) times

$= p^x q^{n-x}$.

But, x successes in n trails can occur in ${}^nc_x$ ways..... which are mutually exclusive and the probability for each of these is $p^x q^{n-x}$. Hence by addition theorem of probability,

$P(X = x) = {}^nc_x p^x q^{n-x}$, x = 0, 1, 2, . . ., n.

Ans to Q No 8: (i) equal, (ii) 2, (iii) 1.5, (iv) $e^{-\lambda(1-e^t)}$ (v) poisson.

Ans to Q No 9: Poisson distribution is also known as probability distribution of rare events, because poisson distribution is a positivaly skewed distribution, i.e probabilities tend to zero for large number of occurrences. For these poisson distribution is also known as probability distribution of rare events.

Ans to Q No 10: Some examples of poisson distribution are,

(i) number of deaths from suicide reported in a particular city.

(ii) number of defictive items in a lot of manufactures items.

(iii) number of printing mistakes in each page of a book.



8.9 MODEL QUESTIONS

Q 1: Define moment generating function and state its properties.

Q 2: Write a note on binomial distribution.

Q 3: Define a binomial distribution distribution and also obtain its mean and variance.

Q 4: Obtain poisson distribution as a limiting case of binomial distribution.

Q 5: If X follows poisson distribution with V(x) = 2, find its p.m.f.

Q 6: Define poisson distribution, state the conditions under which the binomial p.m.f. can be approximated by the poisson p.m.f.

Q 7: Obtain m.g.f. of poisson distribution and hence find its mean.

*** ***** ***

UNIT 9: THEORETICAL PROBABILITY DISTRIBUTIONS (CONTINUOUS VARIABLE)

UNIT STRUCTURE

9.1 Learning Objectives

9.2 Introduction

9.3 Normal Distribution

- 9.3.1 Definition
- 9.3.2 Standard Normal Variate
- 9.3.3 Properties of Normal Distribution
- 9.3.4 Area under Normal Probability Curve
- 9.3.5 Importance of Normal Distribution

9.4 Weak Law of Large Numbers

- 9.4.1 Statement
- 9.4.2 Application of Weak Law of Large Numbers

9.5 Central Limit Theorem

- 9.5.1 Statement
- 9.5.2 Application of Central Limit Theorem

9.6 Let Us Sum Up

9.7 Further Reading

9.8 Answers To Check Your Progress

9.9 Model Questions

9.1 LEARNING OBJECTIVES

After going through this unit, you will be able to-

- discuss about the shape and form of normal distribution, the most important theoretical distribution for continuous variable
- convert a normal variate to a standard normal variate and know its distribution
- understand how normal distribution can be obtained as a limiting case of binomial and poisson distribution
- find the area under the standard normal curve
- explain the important properties of normal distribution regarding its

moments, points of inflexion etc.

- have concept of weak law of large numbers and its application in practical field
- understand the importance of 'Central Limit Theorem' which gives normal distribution the central place in the theory of sampling.

9.2 INTRODUCTION

The binomial and poisson distribution studied in the previous unit are the most useful theoritical distributions for discrete variables, i.e. they relate to the occurrence of distinct events. In order to have a mathematical distribution suitable for dealing with quantities whose magnitudes vary continuously, a continuous distribution is needed. The normal distribution that we are going to discuss in the subsequent section of this unit is the most important theoretical distribution for continuous variable.

So normal distribution is the most widely known and used continuous distribution for data analysis. It is the most important distribution in applied statistics. Many statistical data concerning business and economic problems are displaced in the form of normal distribution. It has got this importance in statistics, because, most of the useful distributions in theoretical statistics, whether discrete or continuous tend to normal distribution when the number of observations are large. For example, normal distribution is an approximation to binomial distribution when the number of independent Bernoulli trials are very large and neither p nor q is very small. Again normal distribution can be obtained as a limiting case of poisson distribution when the mean of the distribution is very large.

Again in section 9.4 of this unit we shall learn about Weak Law of Large Numbers which tells us when a statistic which itself is a random variable will converge in probability to the parameter.

Again it is believed that the main reason for the widespread use of normal distribution in the area of statistical inference is the 'Central Limit Theorem' which we shall discuss in section 9.5 of this unit. According to this theorem the sample mean $\bar{x}$ from a population will approximate a normal distribution when the sample size n is quite large.

9.3 NORMAL DISTRIBUTION

The normal distribution was first discovered in 1733 by De Moivre, a French Mathematican , who obtained normal distribution as a limiting case of binomial distribution and applied it in the game of chance. But it was mainly known to Laplace and Gauss who made reference to it in the beginning of the 19th century.

If we consider a continuous random variable, say, heights of individuals or the yearly rainfall data over a long period of time and plot their probability density functions, we notice one interesting fact : most of those density curves resemble one another in shape, which is bell shapped symmetrical unimodal curve with most densites concentrated near the mode and very little density lying towards the two extremes. Thus most of the frequency, distribution, specially natural phenomena vary in accordance with this curve, called the normal curve, which is an indispensable tool for modern statistics. Let us now define a normal distribution as follows :

9.3.1 Definition

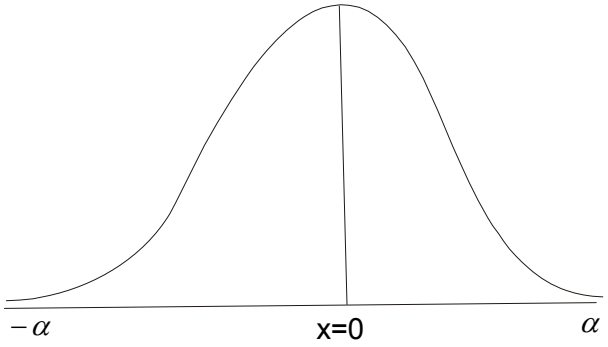
A continuous random variable X is said to have a normal distribution with parameter μ and σ if its density function is given by.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\ e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad ; \quad \begin{matrix} -\infty < X < \infty \\ -\infty < \mu < \infty \\ \sigma > 0 \end{matrix}$$

μ and σ are the mean and standard deviation of this distribution.

Note : A random variable X with mean μ and s.d. σ following normal probability law is expressed as $X \sim N(\mu, \sigma)$.

The probability curve of normal distribution is known as normal curve. This curve is bell shapped, symmetrical about mean μ and two tails axtends to infinity on either side.



Normal curve

9.3.2 Standard Normal Variate

If X is normally distributed with mean μ and standard deviation μ, then the random variable $Z = \frac{X-\mu}{\sigma}$ is called a standard normal variate and the corresponding distribution is called standard normal distribution. If has the density function

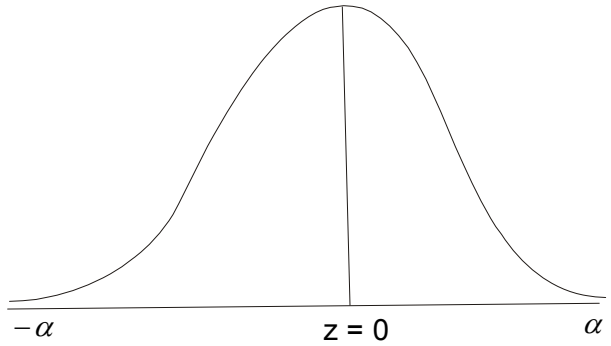
$$\phi(z) = \frac{1}{\sqrt{2\pi}}\ e^{-Z^2/2}\ , -\infty < z < \infty$$

$Z = \frac{X-\mu}{\sigma}$ also follows normal distribution with

E(z) = 0, v(z) = 1.

i.e. Z ~ N (0, 1).

The standard normal curve will look like the ordinary normal curve, only the diffirence is that is symmetrical about z = 0.



Standard normal curve.

9.3.3 Properties of Normal Distribution

A study of the normal distribution reveals the undermentioned significant propertics :

(i) The normal probability curve is bell shapped, symmetrical, unimodal curve. This means that the curve is single peaked and the distribution of frequencies on either side of the mean are same.

(ii) Mean, median and mode co-incide,

i.e.mean=median=mode= μ .

(iii) For normal distribution, skewness = 0

kurtosis = 3

(iv) Sum of independent normal variates is also a normal variate.

(v) All moments of odd order are zero.

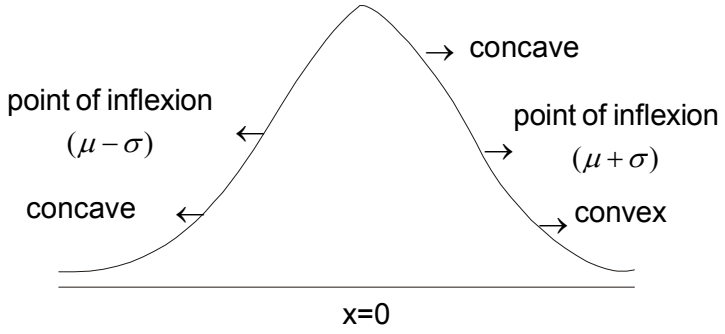
i.e. $\mu_1 = \mu_3 = \mu_5 = \ldots = 0$

And moments of even order are given by the expression

$\mu_{2r} = 1, 3, 5 \ldots . (2r-1)\sigma^{2r}$; r = 1, 2, 3, ...

(i) Quartiles are equidistant from median.

(ii) The points $x = \mu - \sigma$ and $x = \mu + \sigma$ are called the points of inflexion. At these points the curve changes its curvature (i,e changes from convex to concave and concave to convex respectively) and are necessary for finding the shape of the curve.



(viii) The mean deviation is $\frac{4}{5}^{th}$ or more precisely .7979 of the standard diviation. i.e.

$$M.D = \frac{4}{5}\sigma$$

or, M. D = .7979σ .

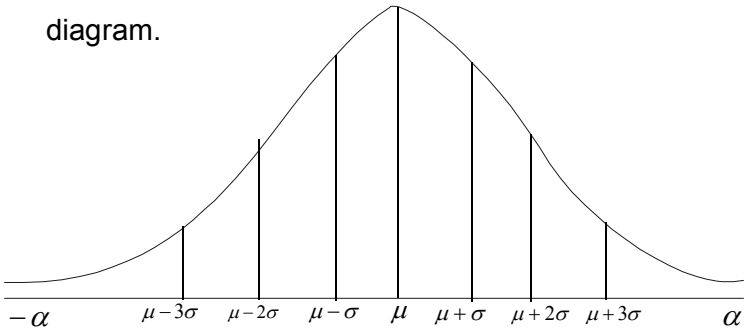
(ix) The area under normal curve is distributed as follows :

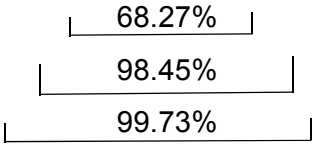
(a) $(\mu \pm 1\sigma)$ covers 68.27% area, 34.135% area will be on either side of the mean.

(b) $(\mu \pm 2\sigma)$ covers 95.45% area.

(c) $(\mu \pm 3\sigma)$ covers 99.73% area.

This distribution of areas can be shown in the following diagram.





Note : Let us now see how the discrete distributions discussed so far like binomial and poisson distributions tend to normal distribution.

(i) If X follows binomial distribution with parameters n and p, then it will tend to normal distribution if the following conditions are fulfilled.

(a) the number of independent Bernoulli trials is very large, i. e. $n \to \infty$.

(b) neither p nor q is very small.

(ii) If X follows poisson distribution with parameter λ, then it will tend to normal distribution if the mean of the distribution is very large, i.e. $\lambda \to \infty$.

9.3.4 Area under Normal Probability Curve

We know that p.d.f. of standard normal variate $Z = \frac{x-\mu}{\sigma}$ is

$$\phi(z) = \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}, \qquad -\infty < z < \infty$$

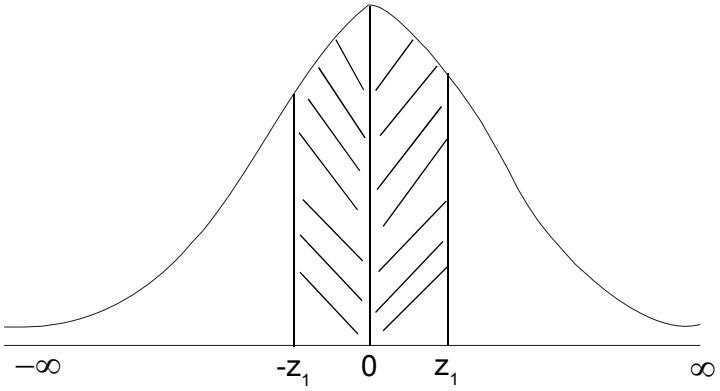
The area under this curve is equal to 1. In any problem, if we are interested in determining the area under normal curve we follow the following steps :

(i) At first we transform the given normal variate X to standard normal variate Z by the transformation $Z = \frac{x-\mu}{\sigma}$.

(ii) We use the table for 'Area under standard normal curve' given at the end of this book. Since in the table areas are given from Z = 0 to a particular positive quantity, so somehow or otherwise we transform the given area to that form.

(iii) Since the normal curve is symmetrical, the probability that Z lies between, say, $-Z_1$ to 0 is equal to the probability that Z lies between 0 to Z_1.

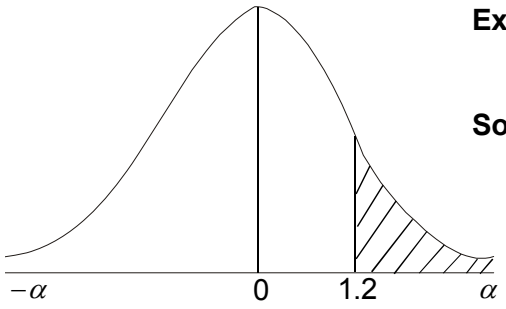
i.e. $P(-z_1 < z < 0) = P(0 < z < Z_1)$



(iv) $P(-\infty < z < -z_1) = 1 - P(-\infty < z < z_1)$

$P(z_1 < z < \infty) = 1 - P(-\infty < z < z_1)$

Now we shall try to find the area under normal probability curve in the following examples.



Example 1: If X ~ N(66, 5), find

(i) $P(X \geq 72)$, (ii) P(65 < x < 70)

Sol^n : (i) $P(X \geq 72)$

$$= P\left(\frac{X-66}{5} \geq \frac{72-66}{5}\right)$$

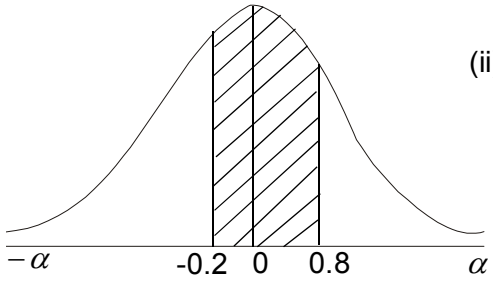
$$= P(z \geq 1.2) \qquad \because Z = \frac{X-\mu}{\sigma}$$

$= P(1.2 < z < \infty)$ [shaded area is the requied area]

$= P(0 < z < \infty) - P(0 < z < 1.2)$

= 0.5 – 0.3849 $\because P(0 < z < \infty) = 0.5$

= 0.115.



(ii) $P(65 < x < 70)$

$$= P\left(\frac{65-66}{5} < \frac{X-66}{5} < \frac{70-66}{5}\right)$$

= P(–0.2 < z < 0.8)

= P(–0.2 < z < 0) + P(0 < z < 0.8)

=P(0 < z < 0.2) + P(0 < z < 0.8) $\because$ P(–4 < z < 0) = P(0 < z < 4)

= 0.0793 + 0.2881 [from table)

= 0.3674

Example 2: How many male workers in a factory have a salary between (i) Rs. 400 and Rs. 680, (ii) more then 720 if mean salary is Rs. 500 and s.d. is Rs. 100, and the salary of the workers assumed to be normally distributed. The number of workers in the family is 10,000.

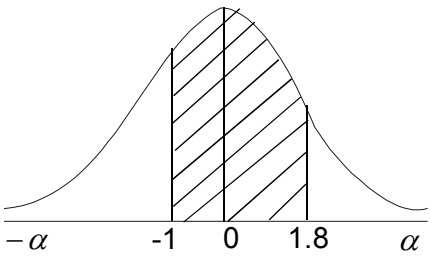
Solution: Let us assume that the random variable X denotes the salary of the workers.

We are given X ~ N (500, 100)

(i) Here we are to find the probability that the salary of the workers is between Rs. 400 and Rs. 680, i.e. we need,

P (400 < X < 680)

$$= P\left(\frac{400-500}{100} < \frac{X-500}{100} < \frac{680-500}{100}\right)$$ [transforming X to Z]



= P(–1 < z < 1.8)

= P (–1 < z < 0) + P (0 < z < 1.8)

= P (0 < z < 1) + P (0 < z < 1.8)

= 0.3413 + 0.4641

= 0.8054

So the number of male workers getting salary between Rs. 400 and Rs. 680 is,

0.8054 x 10,000

= 8054.

(ii) Here we want P(X > 720)

$$= P\left(\frac{X-500}{100} > \frac{720-500}{100}\right)$$

$$= P(z \geq 2.2)$$

$$= P(2.2 < z < \infty)$$

$$= P(0 < z < \infty) - P(0 < z < 2.2)$$

= 0.5–0.4861

= 0.0139.

Hence the number of male workers earning salary more than Rs. 720 is 0.0139 x 10,000

= 139.

Example 3: 1000 light bulbs with a mean life of 120 days are installed in a new factory and their length of life is normally distributed with standard deviation of 20 days.

(i) How many bulbs will expire in less than 100 days.

(ii) If it is decided to replace all the bulbs together, what interval should be allowed between replacements if not more than 10% should expire before replacement.

Solution: Let x devote the mean life of light bulbs, given that $x \approx N(120,20)$.

(i) To find the number of bulbs that expire in less than 100 days, we first find, $P(x < 100)$

$$= P\left(\frac{x-120}{20} \langle \frac{100-120}{20}\right)$$

$$= P(z < -1)$$

$$= P(-\infty < z < -1)$$
$$= P(-1 < z < \infty)$$
$$= P - \{P(-1 < z < 0) + P(0 < z < \infty)\}$$
$$= P - \{P(0 < z < 1) + 0.5\}$$
$$= 1 - 0.8413$$
$$= 0.1587$$

Hence the number of bulbs that should be replaced before 100 days is 1000 x 0.1587
= 159 (approx.)
(ii) The value of z corresponding to an area 0.4 (0.5-0.1) under the normal curve is 1.28.

$$z = \frac{x - \mu}{6}$$
$$\Rightarrow -1.28 = \frac{x - 120}{20}$$
$$\Rightarrow x = 120 - 20 \times 1.28$$
$$= 94$$

Hence the bulbs will have to be replaced after 94 days.

9.3.5 Importance of Normal Distribution

The normal distribution has long occupied a central place in the theory of statistics. Its importance will be clear from the following points.

(i) As n increases, the normal distribution serves as a good approximation to many discrete distributions, such as binomial or poisson distribution.

(ii) One of the main reason behind the extreme use and application of Normal distribution is central Limit Theorem According to this theorem as sample size n increases, the distribution of mean $\bar{x}$, of a random sample taken from practically any population approaches normal distribution.

(iii) The proof of test of significance in sampling theory on the fundemental assumption that the population from which sample has been drawn is normal.

(iv) The normal distribution is used extensively in statistical quality

control in setting up contral limits.

(v) The normal distribution has numerous mathematical properties which make it popular and comparatively easy to manipulate. The normal distribution is reasonably close to many distributions of the humped type. If, therefore, we are ignorant about the exact nature of the distribution, we may assume as a first approximation that the distribution is normal and see what this assumption leads us to.



CHECK YOUR PROGRESS

Q 1: State whether the following statements are true (T) or false (F).

(i) For a normal distribution $\beta_1 = 0, \beta_2 = 3$ (T/F).

(ii) Normal probability curve is a asymmetrical curve. (T/F)

(iii) Normal distribution can be obtained as a limiting form of binomial distribution when $n \to \infty,\ p \to 0$ (T/F)

Q 2: Fill in the blanks.

(i) The probability distribution for which mean = median = mode is known as ..... distribution. (normal/binomial/ poisson).

(ii) If X ~ N (μ, σ), then $P\{\mu - 2\sigma < x < \mu + 2\sigma\}$ is ..... (.9545/.9973)

(iii) The points of inflexion of normal distribution are ..... $(\mu \pm \sigma / \mu \pm 2\sigma)$.

Q 3: Discuss briefly about the shape of the normal curve.

Q 4: Write four properties of normal distribution.

Q 5: There are six hundred students in the post graduate classes of a university and the probability for any student to need a copy of a particular book from the university library on any day is 0.05. How many copiess of that particular book should be kept in the university library so that the probability may be greater than 0.9 that none of the students needing a copy from the library has to come back disappionted. (use normal approximation to binomial distribution).

9.4 WEAK LAW OF LARGE NUMBERS

Suppose $f(x, \theta)$ be the density function of a random variable X. If θ is known, then the density function will be completely specified and no interference from sample to the population will be necessary to consider a random sample which can tell us something about the unknown parameter θ.

Weak law of large numbers tells us that a positive integer n can be determined such that if a random sample of size n or large is taken from a population with density $f(x, \theta)$, with $E(X) = \theta$, the probability that the sample mean $\bar{x}$ will deviate from θ by a very small quantity can be made as close to 1. We now state the Weak Law of Large Numbers as follows.

9.4.1 Statement

Let $X_1, X_2, ..., X_n$ be a sequence of random variables having expectations $\mu_1, \mu_2, ..., \mu_n$. Further let,

$B_n = \text{var}(x_1 + x_2 + ... + x_n) < \infty$

If $\frac{Bn}{n^2} \to 0$ as $n \to \infty$, then, given any two positive quantitices $\in$ and η, however small may be, we can final an n_0 depending on $\in$ and η such that,

$$P\left[\left|\frac{x_1 + x_2 + ... + x_n}{n} - \frac{\mu_1 + \mu_2 + ... + \mu_n}{n}\right| \leq \in\right] > 1 - \eta$$

for all $n \leq n_0$.

Note : For existence of Weak Law of Large Numbers the following three conditions should be satisfied by the random variable X_i, i = 1, 2, ..., n. They are,

(i) $E(X_i)$ exists and is finite.

(ii) $B_n = V(x_1 + x_2 + ... + x_n)$ exists.

(iii) $\frac{B_n}{n^2} \to 0$ as $n \to \infty$.

The condition (i) is necessary, without which the law cannot be stated. Condition (ii) and (iii) are not necessary. However condition (iii) is a sufficient condition.

9.4.2 Application of Weak Law of Large Numbers

Example 4: Let $\{X_n\}$ is a sequence of mutually independent and identically distributed random variables with mean μ and finite variance. If $S_n = X_1 + X_2 + ... + X_n$, prove that the law of large numbers doesnot hold for the sequence $\{S_n\}$.

Solution: Consider the variables $S_1, S_2, ... S_n$.

Now, $B_n = V(S_1, S_2, ... S_n)$

$= V\{x_1+(x_1+x_2)+(x_1+x_2+ x_3) +... (x_1 + x_2 + ... + x_n)\}$

$= V\{nx_1+(n–1)x_2+(n–2)x_3 +...+ 2x_{n-1} + x_n\}$

$= n^2v(x_1)+(n–1)^2 v(x_2)+(n–2)^2 v(x_2) +...+ 2^2 v(x_{n-1})+ 1^2 v(x_n)$

[Covariance terms vanish as $x_1, x_2, ..., x_n$ are independent r. v's and $v(ax) = a^2 v(x)$]

Let, $v(x_i) = \sigma^2$ for all i, since the variables are identically distributed.

$$\therefore B_n = n^2\sigma^2 + (n-1)^2\sigma^2 + (n-2)^2\sigma^2 + ... + 2^2\sigma^2 + 1^2\sigma^2$$

$$= (1^2 + 2^2 + ... + n^2)\sigma^2$$

$$= \frac{n(n+1)(2n+1)}{6}\sigma^2$$

$$\therefore \lim_{n \to \infty} \frac{B_n}{n^2} = \frac{(n+1)(2n+1)}{6n}\sigma^2 \to$$

Hence Law of Large Numbers doesnot hold for the sequence $\{S_n\}$.

9.5 CENTRAL LIMIT THEOREM

A central limit theorem asserts that under quite general conditions, the standardized sum of independent normal variates has an approximately normal distribution. It occupies a place of honour in probability theory acquired by the fruitfull role which it played and still plays in the development of theory applications. It is the central limit theorem which gives the normal distribution it's central place in the theory of sampling, since many important problems can be solved by this single pattern of sampling variability. As a result the work on statistical inference is made simpler. Now we want to state central limit theorem as a generalization of laws of large numbers as follows :

9.5.1 Statement

If $x_1, x_2, ..., x_n$ are n independent random variables with $E(x_i) = \mu_i$ and $v(x_i) = \sigma_1^2$, i = 1, 2, ..., n, then it can be shown that under certain general conditions the random variable $S_n = x_1 + x_2 + ... + x_n$ is asymptotically normal with mean μ and variance σ^2, where,

$$\mu = \sum_{i=1}^{n} \mu_i \qquad \text{and} \quad \sigma^2 = \sum_{i=1}^{n} \sigma_i^2$$

Note : This theorem was first stated by Laplace in 1812 and a rigorous proof under certain general conditions was given by Leapounoff in 1901. Below we shall consider some particular cases of this general central limit theorem.

De- Moivre Laplace Central Limit Theorem :

Statement : If $X_i = \begin{cases} 1, & with\ probability\ p \\ 0, & with\ probability\ q \end{cases}$,

then the distribution of $S_n = x_1 + x_2 + ... + x_n$, where x_i's (i = 1, 2, ...n) are independent, is asymptotically normal as $n \to \infty$.

Lindeberg - Levy Central Limit Theorem :

Statement : If $x_1, x_2, ..., x_n$ are independently and identically distributed random variables with

$$E(x_i) = \mu$$

$$V(x_i) = \sigma^2 \qquad i = 1, 2, ..., n.$$

then the sum $S_n = x_1 + x_2 + ... + x_n$ is asymptotically normal with mean $n\mu$ and variance $n\sigma^2$.

Note : CLT can be stated in another form as follows :

If $x_1, x_2, ..., x_n$ are i. i .d with mean μ and variance σ^2 (finite) and $S_n = x_1 + x_2 + ... + x_n$, then

$$\lim_{n \to \infty} P\left[a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} \le b \right] = \phi(b) - \phi(a) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{z^2}{2}} dz .$$

9.5.2 Application of Central Limit Theorem

Below are given some applications of CLT in commonly used distribution of random variables.

1. If $x_1, x_2, ..., x_n$ are i i d binomial variates with parameters (k, p), 0< p < 1, then,

$S_n = x_1 + x_2 + ... + x_n$. $\sim N(nkp, nkpq)$ as $n \to \infty$

Proof : Now, $E(S_n) = \sum_{i=1}^{n} E(x_i) = n\,kp \quad \because E(x_i) = k\,p$

$$V(S_n) = \sum_{i=1}^{n} v(x_i) = n\,kp\,q \qquad \because V(x_i) = k\,p\,q$$

Hence by CLT we have,

$$\lim_{n \to \infty} P\left\{ a \le \frac{S_n - E(S_n)}{\sqrt{V(S_n)}} \le b \right\} = \phi(b) - \phi(a)$$

$$\Rightarrow \lim_{n \to \infty} P\left\{ a \le \frac{S_n - n\,kp}{\sqrt{nkpq}} \le b \right\} = \phi(b) - \phi(a)$$



CHECK YOUR PROGRESS

Q 6: State true or false.

(i) The law of large number tells us that a statistic will converse to a parameter in some sense. (T/F).

(ii) The central limit theorem are named so because they are central to the distribution theory necessary for statistical inference. (T/F)

Q 7: Let x_i assume values i and –i with equal probability show that WLLN cannot be applied to independent variables $x_1, x_2, ..., x_n$.



9.6 LET US SUM UP

- Normal distribution is the most widely known and used continuous distribution for data analysis.

- p.d.f. of normal distribution is,

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\alpha<x<\alpha \\ -\alpha<\mu<\alpha \\ \sigma>0$$

- μ and σ are the parameters of this distribution, which are also the mean and standard deviation of this distribution.
- Mean, median and mode occur at the same point, i.e mean = median = mode.
- Quartils are equidistance from median.
- It is a symmetrical distribution.
- skewness = 0, kurtosis = 3.
- Binomial distribution tends to Normal distribution if,
  (i) n, the number of independent Bernoulli trials is very large.
  (ii) neither p nor q is very small.
- Poisson distribution with parameter λ tends to normal distribution as $\lambda \to \infty$.
- To find the area under normal curve we have to transform the random variable x to standard normal variate z.
- Central limit theorem establishes conditions under which sum of independent random variables are asymptolically normally distributed.



9.7 FURTHER READING

1) Gupta, S. C. Kapoor, V.K., Fundementals of Mathematical statistics (2002). Sultan Chand and Sons.

2) Goon, A. M, Gupta, M. P. and Dasgupta, B, Fundementals of statistics, Vol I, Seventh Revised Edition (2002). The World Press Pvt. Ltd, Kolkata.

3) Das, K.K and Bhattacharjee, D; A Treatise on statistical Inference and Distribution (2008). Asian Books Private Limited.

4) Kakoty, S. C.; Mathematical Statistics : Theory and applications (2003); Kasturba Prakashan, Dibrugarh, Assam.

5) Saxena, H.C.; Probability and its application (1994); S. Chand and Company, New Delhi.

9.8 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: (i) True, (ii) False, (iii) False.

Ans to Q No 2: (i) Normal, (ii) .9545, (iii) $\mu \pm \sigma$.

Ans to Q No 3: The probability curve of Normal distribution is known as normal curve. The curve is symmetrical about its mean μ. It is bell shapped and extends to infinity on either side. The height of the normal curve is at its maximum at the mean. That is there is only one maximum point of the curve and the height declines as we go on either direction from mean. The curve is asymptotic to the base.

Ans to Q No 4: The four important properties of normal distribution are,

(i) the normal curve is bell shapped and symmetrical about mean.

(ii) for normal distribution, mean = median = mode.

(iii) skewness = 0, kurtosis = 3

(iv) the linear combination of independent normal variates is also a normal variate.

Ans to Q No 5: Let x devote the number of copies of the particular book.

$\therefore$ p(x) = ${}^{n}C_{x}$ $p^x q^{n-x}$, x = 0, 1, 2, ... n.

We are given, n = 600, p = 0.05.

$\therefore \mu = np$ [$\because$ mean of benimal dist^n is np]

= 600 x 0.05 = 3

varianes = $\sigma^2 = npq = 600 \times 0.05 \times 0.95 = 28.8$

$\therefore \sigma = 5.3$

We need x = x, such that

P(X < x) > 0.90

$\Rightarrow P(Z < z) > 0.90$ where $z = \dfrac{x - 30}{5.3}$

$\Rightarrow P(0 < Z < z) > 0.90$

$\therefore$ z > 1.28 from normal probability table.

$$\Rightarrow \frac{x-30}{5.3} > 1.28$$

$$\Rightarrow x > 30 + 5.3 \times 1.28$$

$$\Rightarrow x > 36.784 \cong 37$$

Hence the university library should keep at least 37 copies of that book.

Ans to Q No 6: (i) True, (ii) True.

Ans to Q No 7: Given that, $P(x_i = i) = p$, $P(x_i = -i) = p$

$\therefore$ $E(x_i) = i.p + (-i).P = 0$

$E(x_i^2) = i^2.p + (-i)^2.p = 2\, i^2 p.$

$V(x_i) = E(x_i^2) - \{E(x_i)\}^2 = 2\, i^2 p.$

$\therefore$ $B_n = v(x_1 + x_2 + ... + x_n)$

$= V(x_1) + v(x_2) + ... + v(x_n)$

$= 2.1^2 p + 2.2^2 p + ... + 2.n^2 p$

$= 2p(1^2 + 2^2 + ... + n^2)$

$$= 2p\frac{n(n+1)(2n+1)}{6}$$

$$= \frac{np(n+1)(2n+1)}{3}$$

Now, $\lim_{n \to \infty} \frac{B_n}{n^2} \to \infty$

$\therefore$ Weak Law of Large Numbers can not be applied in this case.



9.9 MODEL QUESTIONS

Q 1: Under what conditions binomial distribution can be approximated to normal distribution.

Q 2: Define normal probability distribution. If the mean of a normal distribution is μ and variance is σ^2, what are its median, mode. β_1 and β_2.

Q 3: Write a note on the salient features of a normal distribution.

Q 4: Define a standard normal variate and obtain its m.g.f.

Q 5: If x follows normal distribution with mean μ and s.d. σ, then find the distribution of Y = a + bX.

Q 6: The monthly income of a group of 10,000 persons found to be normally distributed with mean Rs. 750 and s.d. Rs. 50. Show that of this group, about 95% had income exceeding Rs. 668 and only 5% had income exceeding Rs. 832.

Q 7: State Weak Law of Large Numbers.

Q 8: Let $\{X_n\}$ be a sequence of mutually independent random variables such that $P\{X_n = \pm 2^k\} = \frac{1}{2}$, Does WLLN hold.

*** ***** ***

UNIT 10: SAMPLING AND POPULATION

UNIT STRUCTURE

10.1 Learning Objectives

10.2 Introduction

10.3 Concept of Population and Sample

- 10.3.1 Basic Principles of Sample Survey
- 10.3.2 Advantages of Sampling over Complete Enumeration
- 10.3.3 Different Steps in a Sample Survey
- 10.3.4 Errors in Surveys (Non-Sampling and Sampling Errors)
- 10.3.5 Types of Sampling
- 10.3.6 Selection of Random Samples

10.4 Simple Random Sampling

- 10.4.1 Estimate of Population Mean and Population Total
- 10.4.2 Illustrated Examples

10.5 Stratified Random Sampling

- 10.5.1 Advantage and Disadvantages of Stratified Random Sampling
- 10.5.2 Allocation in Stratified Random Sampling

10.6 Systematic Random Sampling

- 10.6.1 Linear Systematic Sampling
- 10.6.2 Circular Systematic Sampling
- 10.6.3 Advantages and Disadvantages of Systematic Sampling
- 10.6.4 Uses of Systematic Sampling

10.7 Cluster Sampling

- 10.7.1 Advantages and Disadvanges of Cluster Sampling
- 10.7.2 Optimum Cluster size

10.8 Let Us Sum Up

10.9 Further Reading

10.10 Answers To Check Your Progress

10.11 Model Questions

10.1 LEARNING OBJECTIVES

After going through this unit, you will be able to-

- distinguish between a sample and population
- state the difference between parameter and a statistic
- understand why sample surveys are adopted instead of complete enumeration in most cases
- acquire knowledge about the types of errors that one encounters while adopting sample surveys
- distinguish between random sampling and purposive sampling
- draw samples using simple random sampling technique
- discuss the advantages and disadvantages of using stratified random sampling from a heteregeneous population
- explain the process of linear systematic and circular systematic sampling and their relative merits and demerits
- evaluate the advantage of drawing cluster sampling and how to decide optimum cluster size.

10.2 INTRODUCTION

The use of sampling in making inference about a population is a practice that we are doing as a part of our everyday's life from times unknown. When we have to make an inference about a large lot and we are not in a position to examine each and every individual of the lot, we examine only a few members of the lot and make decission about the whole lot. For example a person wanting to purchase some rice will examine a handful of rice from a bagful of rice and on that basis make decisions about the whole bag of rice. Similarly in industry quality of a product is assured through sampling, the public opinion is ascertanied through opinion polls based on sampling. In all these cases inference is drawn about the whole from the knowledge of a part.

So what we see is that most of the enquiries that we make in practice are sampling enquires, sampling is inevitable because our resources are always limited, be it money or man hours. It is also preferred because it has

practical conveniencs and there are many situation where complete count is impossible.

In this unit we shall discuss how to use different sampling techniques to draw our samples according to the type of population and purpose of study. In section 10.3 we shall learn about basic concepts of sampling, its benifits, errors etc. In the subsequent, chapters different sampling techniquies like simple random sampling, stratified random sampling, systematic sampling and cluster samplings are discussed along with their relative advantages and disadvantages.

10.3 CONCEPT OF POPULATION AND SAMPLE

Population : The dictionary meaning of the word population is the number of people residing in a geographical area. But in statistics the word population is used in much wider sense. Thus population may be defined as a group of individuals with respect to certain characteristic under study. For example if we want to measure the average height of students of a college, then the students of that college will constitute a population. Population may be finite or infinite.

Sample : The part of the population which is selected from it in order to study its characteristic is called a sample. For example if we want to measure the standard of living of the people in a town of population 10,000, there if we estimate this characterestic from a properly selected 500 members of the population, then these 500 members of the town will constitute the sample.

Complete enumeration or census : It refers to the enumeration of each and every unit of the population. We all know about the census the government of India conducts in every ten years.

Sample survey : When we enumerate units belonging to the sample only, it is sample survey.

Parameter : The basic objective of any sample survey is to know the characteristic of the sample by examining only a part of it, by studying the characteristic of the population. So the term parameter is used to denote the characteristic of the population. For example population mean, population

variances are parameters. Parametrs are constant quantities.

Statistic : Values obtained from the study of the sample are known as statistic, sample mean, sample variance are the examples of statistic. statistics are variable quantities.

Sampling unit : The individuals that are to be sampled from the population to form the sample is called the sampling units. For example if we want to measure the average income of the families in a city, then each household may be regarded as sampling unit, but if we want to know the average income of individuals in the same city, then the individuals will be the sampling units.

Sampling frame : Sampling frame is the complete list of sampling units. Such sampling frames are difficult to construct and even difficult to keep it update. Simplest example of sampling frame is the voters' list.

Sampling error : It is the difference between the parameter and its estimate from the random sample, It is ineviable because only a part of the population is selected.

Unbiased estimate : Let $T = t(x_1, x_2, ..., x_n)$ be a statistic for the sample values $x_1, x_2, ..., x_n$. This statistic is called an unbiased estimate of the parameter θ, if E (T) = θ.

Sampling variances : The variance of a statistic is called the sampling variance.

Standard error : The positive square root of sampling variance is the standard error.

10.3.1 Basic Principles of Sample Survey

There are two important laws on which the theory of sampling is based. They are principles of validity and optimization.

Principle of validity means that the sampling design should be such that valid estimates and tests about the parameter can be obtained.

Again optimization means either maximization or minimization. According to this principle such sampling design should

be preferred which gives maximum efficiency for a given cost of survey or the minimum cost for a specified level of efficiency.

10.3.2 Advantages of Sampling over Complete Enumeration

In most cases sample survey is preferred over complete enumeration because of the following conditions :

- In sample survey since only a part of the population is selected, there is reduction of cost either in terms of money or in terms of man hours. Although the cost per individual may be larger in sample survey, the total cost is yet expected to be smaller. In most of our cases resources are limited and the result of the survey is required within a specifeed times, it is always imperative to adopt sample survey.
- In certain types of enquirey highly trained personnel or specialised equipments must be used for collection of data. This is not possible in census method because a large group of investigators are to be employed and it will be highly expensive to train such a large group. So for more accurate results sampling is the only way out.
- In sample survey sampling errors cannot be avoided. On the otherhand though sampling error is absent in complete enumeration, large magnitude of non sampling errors crop in, whose magnitude can not be precisely determined and they greatly influence the accuracy of results. However sampling errors can be theoritically calculated and hence its effect on the accuracy of results can be minimized.
- If the enumeration is by its nature destructive, then complete enumeration is not feasible. For example if we want to examine the quality of sweets in a sweet shop, we have to recourse to a sample. Also in case of infinite or hypothetical population sampling is the only course available.

10.3.3 Different Steps in a Sample Survey

Conducting a large scale sample survey involves three main stages : (a) Planning stage, (b) Execution stage, (c) Analysis and reporting stage.

The planning stage consists of the following steps :

- **Defining the objectives :** The objectives of the survey should be clearly stated. The planner should also take into account the available resources, the time limit within which the study should be finished, the degree of accuracy required etc.
- **Defining the population :** The population along with geographical and demographical boundaries must be clearly defined so that no ambiguity arise later on regarding the coverage of the study.
- **Determination of the data to be collected :** This should be done in accordance with the objective of the survey. Once the decision is made about the nature of the data to be collected, one must prepare the questionnaire or a schedue for collection of data.
- **Method of collection of data :** Decision is to be made whether the interview method or the mailed questionnaire method is to be adopted. Though the second method is less costly there is large scope for non-response and is praticable only among the educated masses.
- **Choices of sampling unit :** The sampling unit is the ultimate unit to be sampled for the purpose of the survey. For example depending on the requirement of the survey in a socio-economic survey the ultimate sampling unit may be a family or a member of a family.
- **Getting a sampling frame :** Once the sampling unit is decided, the next step is to see whether the sampling frames of all the units of the population is available. An example of sampling frame can be the list of households from population census of

a region.

- Designing the survey – This is the most important part of planning stage. At this stage decision is made about the type of sampling is to be used and about the need of pilot survey.
- Training of personnel – The investigators and the supervisor should be properly trained before the survey is undertaken.

The **Execution Stage** involves identification of the sampled individual in the field and filling up questionnaire.

The first step in **Analysis and Reporting stage** is to edit completed questionnaires deleting data that are obviousely erroneous. The next step is to tabulate the data and then these tables are utilised for deriving necessary estimatis of population characteristics.

The next step is report writing. The report should incorporate a detailed statement regarding all the stages of the survey and all the statistical informations collected in a neat tabular form. It is necessary to report the amount of error to be expected in the estimates.

At the completion of the survey, the informations should be properly stored for possible use in designing future surveys.

10.3.4 Errors in Surveys (Non-Sampling and Sampling Errors)

To appreciate the need for sample surveys, it is necessary to understand the role of sampling and non-sampling errors in sample survey and complete enumeration. The errors arising due to drawing inference about the population on the basis of sample is known as sampling errors, clearly sampling error is non-existent in complete enumeration, since the whole population is enumerated. Howerver some other kind of error crop-in, at the stage of ascertainment and, precesing of data, which are termed as non-sampling errors and are common in both complete enumeration and sample surveys. We shall discuss about these two types of errors one by one.

Non sampling errors : As discussed above non-sampling errors are present in both complete enumeration as well as sample surveys. Below we shall study different types of non-sampling errors.

- **Respons errors:** These errors have their origin in the responses or informations provided by the respondents. For example wrong answers arising from pride, called error due to prestige, by virtue of which one may overstate ones educational qualification, or occupation or understate one's age. Sometimes one may provide wrong answers for the protection of self interest. For example people generally tend to make an understatement of their incomes and overstatement of their expenses. Sometimes respondent may misunderstands a particular question and hence may provide wrong answers un willingly.
- **Errors arising from non-response:** Often at the times of collection of data some respondents are not found at home even after repeated calls, or he or she simply refuses to furnish information. As a result of these non-responses the effective simple size decreases and certain characteristic of the population gets excluded, which gives rise to error.
- **Observational error :** This type of error occurs when the variate value is obtained by observation, because of the influence of somes psychological factors. For example in case of eye condition of crop condition, the crop reporters invariably understimate the result.
- **Interviewer error :**Sometimes the suggestions from the interviewer may influence the beliefs and prejudices of the respondents in interpreting some questions and thereby leading to some error.

Sampling Errors :

These errors have their origin in sampling and are absent in complete enumeration. Some errors of this type are mentioned below.

- **Error due to defective sampling technique :** Samples can

generally be drawn in two ways– (a) Random sampling and (b) Purposive sampling.

Random sampling is free from any type of errors because of the random nature of sampling. But in case of purposive sampling the sampler deliberately tries to select a representative sample to establish his own belief, which obviousey gives rise to errors. For example in a socio-economic survey of the city of Guwahati, the sampler may select only the people belonging to upper class if the investigator wants to establish that the people of the city of Guwahati are very rich.

- **Error due to substitution :** Investigators often substitute one convenient member of the population when difficulties arise in enumerating another, which may lead to errors as the characteristics of the two persons may be quite different.
- **Error due to wrong choice of statistic :** Such errors may arise when the investigator doesnot use appropriate statistic for the estimation of a particular parameter of the population. For example, we know that sample mean square is an unbiased estimate of the population variance, but one may use sample variance in place of sample mean square, leading to error.

10.3.5 Types of Sampling

Whenever we decide to take sample from a population it is necessary to choose some method in which sampling is to be done. There are various methods of choosing a sample from a population, depending on the nature of the data and the purpose of enquiry. However these various methods of sampling can be grouped under two broad heads : Probability sampling (also known as random sampling) and non-probability (non random) sampling.

Probability sampling : Probability sampling methods are these in which every item in the universe has a known chance or probability of being chosen for the sample. This implies that the selection of the sample item is independent of the person making

the study, i.e the items are chosen strictly at random.

Non Probability sampling : Non-probability sampling, also called 'Purposive' or 'judgement' sampling is the process where the sample is selected on the basis of experenice or according to the conveneince of the user. Hence unlike random sampling, here, the probability of selection of some of the units of the population are very high, and some are less. This method is not recommended for general use. However in case of experined investigator judgement sampling may give valuable results.



CHECK YOUR PROGRESS

Q 1: Fill in the blanks.

(i) If all the units of the population are enumerated then it is called ..... (sample/population)

(ii) A population consisting of unlimited number of units is called a ..... population. (finite/infinits)

Q 2: State whether the following statements are true or false.

(i) Sampling errors are present both in complete enumeration and sample survey. (T/F).

(ii) Judgement sampling is not recommended for general use. (T/F)

Q 3: In what situation sampling is inevitable.

Q 4: Mention in brief the objective of sampling.

Q 5: Distinguish between parameter and statistic.

10.3.6 Selection of Random Samples

Random samples are mainly selected either by (i) Lottery Method or (ii) Random number table method.

Lottery Method : In this method all the units of the population are numbered and the numbers are written on some cards which are identical in shape, size and colour. These cards are then placed in a jar and shuffled properly. A blind-fold selection is made from these cards untill the desired sample size is acheived. However shuffling

is necessary after each draw. This is a very reliable method of selection of random sample. But when the population is large this method is not practical.

Random number method : This is the most widely used method of selection of random samples. Here at first we number the population. Then we select at random any page of a random number table and pick up numbers in any row or column at random. Items corresponding to numbers thus selected constitute a random sample.

Note : Random number tables are so constructed that each of the digits 0, 1, 2, ..., 9 occurs independently and approximately with same frequency. Some commonly used such tables are 'Teppets Random Number Table'. 'Fisher and Yates Table' etc.

10.4 SIMPLE RANDOM SAMPLING

The simplest and the most commonly used type of probability sampling is simple random sampling. In this sampling each unit of the population has the same probability of being included in the sample. Simple random sampling may be of two types – simple random sampling with replacement and simple random sampling without replacement, as discussed below :

Simple random sampling with replacement (SRSWR) :

In this type of sampling the members of the sample are drawn one by one from the population and after each draw and the selected number is returned back to the population. As a result any member of the population may appear more than once in the sample. So if we want to draw a sample of size n from a population of size N, then each member of the population has the same probability $\frac{1}{N}$ of being selected and the total number of possible samples is N^n.

Simple random sampling without replacement (SRSWOR) :

In SRSWOR the sample units are drawn one by one and the members selected at any draw are not returned back to the population.

Obviously in this case no member of the population can appear more than once in the sample, i.e. the sample units are distinct. If we want to draw a sample of size n from a population of size N, the chance of selecting any unit at the first draw is $\frac{1}{N}$, at the 2nd draw it is $\frac{1}{N-1}$, at the third draw $\frac{1}{N-2}$ and so on. In SRSWOR the total number of possible samples are N_{c_n}.

10.4.1 Estimate of Population Mean and Population Total

Before going to find the estimate of population mean and population total we shall have to know some important notations as given below.

The N units of the population are denoted by X_1, X_2, ..., X_N

The n units of the sample are denoted by $x_1, x_2, ..., x_n$.

$\mu = \frac{1}{N}\sum_{i=1}^{n} X_i$ is the population mean.

$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n} (X_i - \mu)^2$ is the population variance.

$S^2 = \frac{1}{N-1}\sum_{i=1}^{n} (X_i - \mu)^2$ is the population mean square.

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean.

$s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$ is the sample variance.

$s'^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$ is the sample mean square.

Using the above notations we can state a very important theorem as stated below.

Theorem. 10.4.1

Statement : In simple random sampling sample mean is an unbiased estimate of the population mean.

In symbol, we can write,

$E(\bar{x}) = \mu$

Again we know that population total is $\sum_{i=1}^{n} X_i = N\mu$

$\therefore$ Estimate of population total is $N\bar{x}$

i.e. $E(N\bar{x}) = N\mu$

Some useful Results :

(i) $\frac{n}{N}$ is the sampling fraction and is considered negligible if its valus is less than 0.05.

(ii) $\frac{N-n}{N-1}$ is called the finite population correction factor and may be ignored if N is vary large compared to n.

(iii) In SRSWR, the sampling variance of estimate of population mean is given by, $V(\bar{x}) = \frac{\sigma^2}{n}$.

(iv) In SRSWOR, the sampling variance of estimate of population mean is given by, $V(\bar{x}) = \frac{N-n}{N-1}\frac{\sigma^2}{n}$.

(v) $V(\bar{x})_{SRSWOR} < V(\bar{x})_{SRSWR}$

Note : The sampling design having less variance is always more efficient. Hence SRSWOR is more efficient than SRSWR sampling design.

10.4.2 Illustrated Examples

Example 1: Suppose that 4 units of a population are $X_1 = 2$, $X_2 = 4$, $X_3 = 6$, and $X_4 = 8$. Draw all possible random samples of size 2 without replacement and calculate their means, show that $E(\bar{x}) = \mu$.

Sol^n : Here we have N = population size = 4

n = sample size = 2

Total no. of samples drawn by SRSWOR is $4_{c_2} = 6$, as given below:

Serial No.	Sample	Sample mean (x_i)
1	(2. 4)	$\frac{2+4}{2}$ =3
2	(2, 6)	4
3	(2, 8)	5
4	(4, 6)	5
5	(4, 8)	6
6	(6, 8)	7
Total		$\sum_{i=1}^{n} \bar{x}_i = 30$

$$E(\bar{x}) = \frac{1}{{}^4c_2} \sum_{i=1}^{n} \bar{x}_i = \frac{30}{6} = 5$$

$$\mu = \frac{X_1 + X_2 + X_3 + X_4}{4} = \frac{2+4+6+8}{2} = 5$$

$\therefore E(\bar{x}) = \mu$.



CHECK YOUR PROGRESS

Q 6: State true or false.

(i) When every unit of the population has the same probability of coming to the sample it is called simple random sampling. (T/F)

(ii) In SRSWOR the same item may appear more than once in the sample.

Q 7: Fill in the blanks.

(i) $V(\bar{x})_{SRSWOR}$ = ...

(ii) In simple random sampling estimate of population total is .....

Q 8: Distinguish between SRSWR and SRSWOR.

10.5 STRATIFIED RANDOM SAMPLING

Till now we have discussed the technique of drawing simple random samples and also obtained estimates of the population parameters. But simple random sampling can give a representative sample only when the

population under consideration is homogeneous with respect to the characteristic under study. But in practice we come across populations which are not homogeneous in nature. The sample technique used for such a population is stratified random sampling which by using available information concerning the population, attempts to design a more efficient sample than obtained by simple random sampling technique. While applying stratified random sampling technique, we follow the steps given below:

(i) The population to be sampled is subdivided into groups called strata which are non-overlapping and collectively exhaustive.

(ii) A simple random sample is then chosen independently from each group, untill the required sample size is achieved.

However, for maximum efficiency the strata should be so formed that the units within a strata are as homogeneous as possible and a marked difference is possible between the strata.

For illustration suppose, Let our heterogeneous population of size N is divided into k strata, so that, there are

N_1 units in the 1^{st} stratum

N_2 units in the 2^{nd} stratum

. . . . . .

N_k units in the k^{th} stratum.

So that $N_1+N_2+.....+N_k = N$.

We now want to draw a sample of size n from this population so that we draw.

n_1 units from the 1^{st} stratum

n_2 units from the 2^{nd} stratum

. . . . . .

n_k units from the k^{th} stratum

such that $n_1+n_2+.....+n_k = n$. Only if the samples from each stratum are drawn using Simple Random Sampling then the method is called stratified random sampling.

10.5.1 Advantages and Disadvantages of Stratified Random Sampling

Advantages :

- This method ensures better representation of the characteristic of the population.
- Division of a big population into small sub-populations can bring administrative convenience. For example the agency conducting the survey may employ different field officers for different parts of the population.
- Sometimes different degree of accuracy is needed for different sections of the population, which may be possible only in stratified random sampling.
- In case of skewed data like income data this method is very effective.

Disadvantages :

- It is important to decide the basis of stratification well in advance, otherwise the results may be misleading.
- Each stratum must have, as far as possible homegeneous items, otherwise the results may not be reliable. If proper stratifiction is not done, the sample may have the effect of bias.
- Because of the likelihood that the stratified sample will be more widely distributed geographically than a random sample, cost per observation may be quite high.

10.5.2 Allocation in Stratified Random Sampling

In stratified random sampling, the allocation of sample to different strata is done by considering three factors, viz.

- the total number of units in the stratum,
- the variability within the stratum,
- the cost in taking observations per sampling unit in the stratum.

A good allocation is one where maximum precision is obtained with minimum resources. In otherwords, allocation of sample size

should be such as to minimize the budget for a given variance or minimize the variance for a fixed budget, so that there is most effective use of available resources.

There are three methods of allocation of sample size to different strata in stratified sampling technique. They are: Proportional allocation, Neyman allocation and Optimum allocation.

- **Proportional allocation :** This allocation was originally proposed by Bowley (1926). This procedure of allocation is very common in practice because of its simiplicity, where the size of the sample from each stratum varies with the size of the stratum. Larger the size of the stratum. Larger is the size of the sample selected from that stratum. In proportionall allocation, one samples each stratum in proportion to its relative weight.
- **Neyman allocation :** This allocation of total sample size to strata is called meximum variance allocation and is due to Neyman (1934). Here the allocation of samples among different strata is proportional to the product of the size and variability of the stratum.
- **Optimum allocation :** In this method of allocation of the sample size in the respective strata with a view to minimize variance for a specified cost of the survey or to minimize the cost for a specified value of the variance.

Illustrated example :

Example 2: The population of three towns are 40,000, 30,000 and 20,000 respectively. A stratified random sample of size 400 is to be drawn from this population using proportional allocation. Determine the sample size salected from each stratum.

Solution: Here, $N_1 = 40,000$, $N_2 = 30,000$, $N_3 = 20,000$

$\therefore N = N_1+N_2+N_3 = 90,000$

sample size = n = 400

Now, n_1 = sample size from 1st stratum

$$= 40{,}000 \times \frac{400}{90{,}000} = 177.7 \approx 178$$

$$n_2 = 40{,}000 \times \frac{400}{90{,}000} = 133$$

$$n_3 = 40{,}000 \times \frac{400}{90{,}000} = 89$$

Obveovity $n_1 + n_2 + n_3 = 400 = n$.



CHECK YOUR PROGRESS

Q 9: State wheter the following statements are true (T) or false (F).

(i) stratified sampling is preferred when the population is homogeneous. (T/F)

(ii) Simple Random sampling method is more efficient than stratified random sampling method (T/F)

10.6 SYSTEMATIC RANDOM SAMPLING

Till now we have discussed sampling techniques where sampling units were selected randomly. The sampling technique we are going to discuss now is also known as quasi-random sampling technique, because in this sampling technique only the first unit is selected at random. Once the interval starting point is determined the remainder of the items selected for the sample are predetermined by the sampling interval.

So a systematic sample is formed by selecting one unit at random and then selecting additional units at evenly spaced intervals untill the desired sample size is achived. This method is popularly used in those cases where a complete list of the population from which sample is to be drawn is available. The list may be prepared in alphabitical, geogrephical, numerical or some other order. The items are then serially numbered. The first item is selected at random, generally by using 'Lottery Method'. Subsequent items are selected by taking every k^{th} unit of the list, where k stands for sampling interval, i.e. $K = \frac{N}{n}$

where, n = sample size,

N = population size

Systematic sampling is simple and fool proof. We shall first discuss the sample selection procedures before going to discuss its advantages and desadvantages.

10.6.1 Linear Systematic Sampling

As discussed above, a common procedure is the linear systematic sampling scheme. We suppose that the N units of the population are linearly ordered from which a sample of size n is to be drawn. Here two cases may arise :

Case 1 : Let N be expressible in the form N = nk, where k is an integer, called the sampling interval.

In this procedure, a random number $i \leq k$ is selected. Here the sample will consists of the unit and every k^{th} unit thereafter. Hence, under this procedure the sample comprises of the units $i, i+k, i+2k, ..., i+(n-1)k$.

The technique will generate k systematic samples with equal probability which may be shown in the following table.

Sample Number			
1	2		k
y_1	y_2		y_k
y_{1+k}	y_{2+k}		y_{2k}
y_{1+2k}	y_{2+2k}		y_{3k}
.	.		.
.	.		.
.	.		.
$y_{1+(n-1)k}$	$y_{2+(n-1)k}$		y_{nk}

Case 2: N is not expressible in the form N = nk

Here the present sampling scheme will give rise to samples of unequal sizes. Here also we follow the same procedure with the

random selection of the 1st unit and selection of every k[th] unit thereafter. Since in this case $N \neq nk$, k is taken as an integer nearest to $\frac{N}{n}$. Under this scheme the sample size is not necessarily n and in some cases it may be less than n.

To make this concept more clear let us consider a population of 12 units from which we are to draw a systamatic sample of size 4.

i.e N = 12, n = 4

In this case N = nk where k = 3. Hence the possible samples will be,

$i\ (\leq k)$	Selected sample
1	(1, 4, 7, 10)
2	(2, 5, 8, 11)
3	(3, 6, 9, 12)

Now, let N = 11, n = 4.

So $N \neq nk$

$$\therefore k = \frac{11}{4} = 2.7 \cong 3$$

So the samples will be.

$i\ (\leq k)$	Selected sample
1	(1, 4, 7, 10)
2	(2, 5, 8, 11)
3	(3, 6, 9)

i.e, when N is not a proper multiple of n all the samples are not of equal size. To overcome this difficulty of varying sample size we may use circular systematic sampling.

10.6.2 Circular Systematic Sampling

So, as discussed above the disadvantage of sample size being different from actual sample size required, when $N \neq nk$, the procedure is slightly modified by which a sample of constant size is always obtained. The procedure consists in selecting a unit by a

random start, from 1 to N and thereafter selecting every k^{th} unit, k being an integer nearest to $\frac{N}{n}$, in a circular manner, untill a sample of n units is obtained.

For illustration, let N = 11, n = 4, so that k = 3 (apprx.) So the possible samples are,

(1, 4, 7, 10), (2, 5, 8, 11), (3, 6, 9, 1), (4, 7, 10, 2), (5, 8, 11, 3), (6, 9, 1, 4), (7, 10, 2, 5), (8, 11, 3, 6), (9, 1, 4, 7), (10, 2, 5, 8), (11, 3, 6, 9).

10.6.3 Advantages and Disadvantages of Systematic Sampling

Advantages :

1) Systematic sampling is operationally more convenient than simple random sampling or stratified random sampling, because it requires less time and less work involved.
2) Systematic sampling may be more efficient than simple random sampling when the list from which the sampling units are drawn is arranged at random.

Disadvantages :

1) Systematic sampling becomes less representative if we are dealing with populations having 'hidden periodicities'.
2) If the population is ordered in a systematic way with respect to the characteristic under study, then it is possible that only certain types of items will be included in the sample, or at least more of certain types of items than others. For instance, in a study of workers wages the list may be such that every tenth worker on the list gets above Rs. 2,000 per month.

10.6.4 Uses of Systematic Sampling

Systematic sampling is operationally convenient and efficient in sampling some natural population like forest areas for estmating volume of timber, areas under different types of covers etc.



CHECK YOUR PROGRESS

Q 10: State whether the following statements area true (T) or false (F).

(i) When the population consists of units arranged at random one would prefer systematic sampling. (T/F)

(ii) When the population size is not a proper multiple of sample size, one should prefer circular systematic sampling rather than ordinary systematic sampling. (T/F)

Q 11: In what situation systematic sampling is preferred over other sampling techniques.

Q 12: Compare systematic sampling with stratified sampling technique.

10.7 CLUSTER SAMPLING

In random sampling we have seen that the population is divided into a finite number of distinct and identifiable elements known as sampling units. The smallest units into which the population is divided is called an element of the population. A group of such elements is called a cluster. So when the sampling unit is a cluster, rather than a single element of the population, the sampling procedure is known as cluster sampling.

Therefore, in cluster sampling the entire area containing the population is divided into some smaller area segments and each element of the population belongs to one and only one segment of the population. This procedure is also called ‘area sampling’.

Generally, in any sampling procedure the main problem lies in identification and location of an element. However once an element has been located, the time taken for surveying a few neighbouring elements is less. That is where the justification for formation of cluster lie. But as clusters are made-up of neigbouring elements, they may have similar characteristies. Hence it is always advisable to form small clusters and the number of clusters should be more.

After dividing the population into clusters, the required number of

clusters are selected by equal or unequal probabilities of selection. Again all elements in selected clusters are enumerated.

10.7.1 Advantages and Disadvantages of Cluster Sampling

Advantages :

- Cluster sampling is less expensive and requires less times compared to other sampling techniques discussed so far. It is so because selection of neighbouring elements is easier, cheaper, faster and operationally more convenient.
- This technique can be used when sampling frame of elements is not available and it is too expensive and time consuming to prepare such a list.
- It is less costly than SRS due to savings of time of journey, identification etc.

Disadvantages :

- In general cluster sampling is less efficient than SRS. However, efficiency of cluster sampling increases with decrease in cluster size and increase in variation within cluster.

Note : Various sampling procedures can be applied to cluster sampling by treating clusters.

10.7.2 Optimum Cluster Size

Efficiency of cluster sampling increases with the increase in the number of clusters. However, the cost of survey also increases with increase in the number of clusters. But in any survey our resources are always limited. That is we must try to minimize cost. Hence it is necessary to determine a balancing point by finding out the optimum cluster size and the number of clusters in the sample so that maximum efficiency can be obtained for a given cost or minimum cost is incurred for a given level of efficiency.



CHECK YOUR PROGRESS

Q 13: State whether the following statements are true or false.

(i) A cluster consists of a number of elimentary units of the population. (T/F)

(ii) Efficiency of cluster sampling increases as cluster size increases. (T/F)

Q 14: In what situation cluster sampling is preferred?



10.8 LET US SUM UP

- Population is a group of individuals with respect to certain characteristic under study.
- Sample is a part of the population selected from it in order to study its characteristic.
- Census or complete enumeration refers to enumeration of each and every unit of the population.
- Sample survey refers to enumeration of each and every unit of the sample.
- Two basic principles of sample survey are validity and optimization.
- Non-sampling errors are present in both census as well as sample survey, whereas sampling errors are present only in sample survey.
- Sampling is broadly classified into two categories– probability sampling and non-probability sampling.
- Stratified random sampling is used when the population under consideration is not homogneous with respect to the characteristic under study.
- stratified random sampling is more efficient than SRS.
- In systematic sampling techneque, only the first unit is selected at random and the rest gets automatically selected according to some pre-designed pattern.
- In systematic sampling when the population size N is a multiple of the sample size n, linear systematic sampling is used and when N is not

a proper multiple of n, circular systematic sampling is to be used to overcome the difficulty of varying sample size.

- In cluster sampling, the population is divided into some non-overlapping clusters and required number of clusters are selected by equal or unequal probabilities of selection.
- Efficiency of cluster sampling increases with the increase in the number of clusters, but at the same time cost also increases. Hence, it is necessary to balance between them so that maximum efficiency can be abtained.



10.9 FURTHER READING

1) Cockran, W. G. Sampling Techniques (1963), John Wilay and Weley Eastern.

2) Goon, A. N, Gupta, M. K, Dasgupta, B. Fundemantals of Statistics (Vol II). (1986). The World Press Private Limited.

3) Singh, Daroga, Chaudhury, F. S. Theory and Analysis of Sample Survey Designs. (1986). New Age International Publishers.

4) Sukhatme, P. V. and Sukhatme, B. V. Sampling Theory of Surveys with Applications (1970). FAO and Asia Publishing House.



10.10 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: (i) Population census, (ii) infinite.

Ans to Q No 2: (i) True, (ii) True, (iii) False.

Ans to Q No 3: Sampling is inevitable under the following conditions.

(i) When the resources for survey, i.e. money and trained man power is limited,

(ii) When there is time constraint,

(iii) When the population is very large,

(iv) When the survey is destructive in nature,

(v) When survey is to be conducted in a very wide area.

Ans to Q No 4: The main objective of the survey is to obtain maximum information about the population at minimum cost, time or labour.

Ans to Q No 5: The difference between parameter and statistic is as follows:

(i) The parameter are population characteristics such as population means, population variance but statistics are sample characteristics such as sample mean, sample variance etc.

(ii) Parameters are constant quantities but statistics are variables.

Ans to Q No 6: (i) True, (ii) False.

Ans to Q No 7: (i) $\frac{N-n}{N-1}\ \frac{\sigma^2}{n}$ (ii) $N\bar{x}$

Ans to Q No 8: If the units of the population are drawn one by one in such a way that the units drawn are returned back to the population after each draw, it is known as SRSWR. In this type of sampling the probability of drawing a unit at each draw remains $\frac{1}{N}$. In this case a unit may appear more than once in the sample.

On the otherhand if the units selected from the population are drawn one by one but are not returned back to the population, it is called SRSWR. Here the probability of drawing any unit at the r^{th} draw is $\frac{1}{N-r+1}$. In this type of sampling the same unit can never be included in more than one sample.

Ans to Q No 9: (i) False, (ii) True.

Ans to Q No 10: (i) True, (ii) False.

Ans to Q No 11: Systematic sampling is preferred when informations to be collected about trees in a forest, houses in a block, entries in a register, which are in serial order.

Ans to Q No 12: Systematic sampling resembles stratified sampling in the sense that the population is divided into strata of size k where one unit is drawn from each stratum. But the fact that the strata should be homogeneous is not followed here.

Again no independent samples are drawn from each group. Also in systematic sampling we have only one sample from the whole

population.

From above discussion it is clear that systematic sampling is quite different from stratified sampling.

Ans to Q No 13: (i) True, (ii) False.

Ans to Q No 14: The cluster sampling is preferred when

(i) the sampling frame is not available or too expensive and time consuming to prepare it,

(ii) the sampling units are distance apart,

(iii) the elimentary units may not be easily identifiable and locatable.



10.11 MODEL QUESTIONS

Q 1: Give a brief account of advantages and disadvantages of sampling.

Q 2: What is simple random sampling? What are its advantages?

Q 3: What is stratified random sampling? Why stratified random sampling is more advantageous than SRS?

Q 4: Distinguish between random sampling and judgement sampling.

Q 5: Distinguish between linear and circular systematic sampling.

Q 6: Describe briefly the different sources of error in sample survey.

Q 7: What are the advantages of sample survey over complete enumeration.

Q 8: What is simple random sampling? Describe two methods of selection of selection of simple random sample.

Q 9: A simple random sample of 30 households was drawn from a city having 14848 households. The number of in the selected households are :

5, 6, 3, 3, 2, 3, 3, 4, 4, 3, 2, 7, 4, 3, 4, 4, 3, 3, 4, 4, 3, 3, 1, 2, 4, 3, 4, 3, 4, 3. Estimate the population total.

Q 10: Write short notes on: (i) Principles of sample survey, (ii) Non-sampling error. (iii) Cluster sampling, (iv) Systematic sampling.

*** ***** ***

UNIT 11: CHARACTERISTICS OF A GOOD ESTIMATOR

UNIT STRUCTURE

11.1 Learning Objectives

11.2 Introduction

11.3 Estimators and Estimates

11.4 Types of Estimation

11.5 Criteria for a Good Estimator

- 11.5.1 Unbiasedness
- 11.5.2 Consistency
- 11.5.3 Efficiency

11.6 Let Us Sum Up

11.7 Further Reading

11.8 Answers To Check Your Progress

11.9 Model Questions

11.1 LEARNING OBJECTIVES

After going through this unit, you will be able to-

- discuss how the theory of estimation is used to draw inference about population characteristies from the information provided by the sample
- distinguish an estimator from an estimate
- explain the difference between point estimation and interval estimation
- select an appropriate statstic to serve as a best estimator of a population parameter.
- have knowledge about the propertices that a good estimator should possess.
- put forward the concept of unbiasedness, consistency and efficiency of an estimator.

11.2 INTRODUCTION

This unit deals with the basic concepts of the theory of estimation. We have already discussed about the terms like population, sample, parameter and statistic etc in unit a of this book. So in the light of these knowledge we may define estimation as the technique of coming to a conclusion, regarding the magnitude of a parameter based on the information provided in the sample. As parameters reflect the characteristis of the population, whose values are unknown, our only way to estimate their values is to depend on the random sample from the same population. Obviously greater the size of the sample better will be the estimate of the parameter.

If we consider the case of a normal distribution with parameters μ and σ, whose values are unkonwn, there will always be an infinite number of functions of the sample values, called statistic, which we may propose as the estimates of the parameters. But the best estimate will be the one which will be the nearest to the true value of the parameter to be estimated.

In the subsequant sections of the unit, we shall discuss about a few related terms and concepts like estimators and estimates, types of estimation to develope a clear understanding of the topic, Also we are going to learn the concept of unbiasedness, consistency and efficiency with the help of examples.

11.3 ESTIMATORS AND ESTIMATES

In the method of estimation we are interested in finding an estimated formula of the estimate of the parameter, which is called the **'estimator'**. Again the numerical value of the parameter obained on the basis of the random sample using that estimator is called the **'estimate'**.

Thus estimator is the sample statistic used for obtaining an estimate of the parameter being estimated. For example, we know that the sample mean $\bar{x}$ is used to estimate mean μ, then $\bar{x}$ is an estimator. Again, an estimate refers to the actual sample result used to represent μ. If from a sample of size n = 36, the avarage age is 20 years, these sample mean $\bar{x}$ = 20 years is the estimate.

11.4 TYPES OF ESTIMATION

There are two types of estimation for a population parameter. They are,

(i) Point estimation

and (ii) Interval estimation.

In the first type of estimation, in the light of the sample, one may put forward a single value (i, e a point) as the estimate of the parameter. That is why it is called point estimation.

Sometimes instead of finding a point estimator, one may be interested in finding an interval within which the true value of the parameter lies. The method of construction of such an interval is known as interval estimation. To understand the difference betwen two types of estimation, think of an enumerator collecting data on the number of members per family, i.e, average family size in Assam. On the basis of his sample of say, 1000 familices, the enumerator may find the average family size may be taken as 4.2. This is point estimation.

On the otherhand, again, on the basis of the same sample, he may find that the average family size supposed to lie between 3.9 to 4.6, so the interval 3.9 to 4.6 is the interval estimate.

11.5 CRITERIA FOR A GOOD ESTIMATOR

It follows that an estimator must possess some desirable properties to be a good estimator. Such properties must meet the criterion that a good estimator is one whose sampling distribution is clustered around the true value of the parameter being estimated, than that of any other estimator, so a statistic can be regarded as a good estimator of a parameter if it satisfies the following properties.

(i) Unbiasedness

(ii) Consistency

(iii) Efficiency.

(iv) Sufficiency

We shall discuss about these properties one by one.

11.5.1 Unbiasedness

It is very likely that, the value of a statistic calculated from a given sample will be above or below the actual values of the population parameter of interest, because of the presence of sampling error. So what is desirable is that, the expected value or mean of the possible values of a statistic estimated from over all possible samples should be equal to the population parameter being estimated. If this is true then the statistic is called are unbiased estimator of the population parameter.

So the sample statistic $\bar{x}$ will be an unbiased estimator of the population parameter θ, if,

$$E(\bar{x}) = \theta$$

where, $E(\bar{x})$ is the mean of the sample statistic $\bar{x}$.

In the sampling distribution of sample mean and sample proportion, we get,

$E(\bar{x}) = \mu$ and $E(\bar{p}) = P$

Therefore, $\bar{x}$ and $\bar{p}$ are unbiased estimators of the corresponding population parameters μ and P respectively. Similarly, the difference $\bar{x}_1 - \bar{x}_2$, i.e. the difference between two sample means, $\overline{p_1} - \overline{p_2}$, the difference between two sample proportions are unbiased estimators of population parameters $\mu_1 - \mu_2$ and $P_1 - P_2$ respectivaly.

Again, if, $E(\bar{x}) \neq \theta$, then $\bar{x}$ is called a biased estimator of the parameter θ, and the amount of bias is, $E(\bar{x}) - \theta$,

Again, the sample mean square, not the sample variance is an unbiased estimators of the population variance σ^2.

i.e., $E(s'^2) = \sigma^2$

Where, $s'^2 = \frac{1}{n-1}\sum_{i-1}^{n}(x_i - \bar{x})^2$ is called the sample mean square, However, for large n, the sample variance, i. $s^2 = \frac{1}{n}\sum_{i-1}^{n}(x_i - \bar{x})^2$

becomes an unbiased estimator of the population variance σ^2.

It may be however mentioned that many biased estimators tend to become unbiased estimators asymptotically, because the bias vanishes as n increases.

An important point to be mentioned here is that an unbiased estimator may not be unique. To explain this let us consider the case of a poisson distribution with parameter λ.

As we know for Poisson distribution mean = λ = variance

Now, as discussed earlier, the sample mean is an unbiased estimator of the population mean.

i.e. $E(\bar{x}) = \lambda$

And sample mean square is an unbiased estimator of the population variance.

i.e. $E(s'^2) = \lambda$

So both sample mean and sample mean square are unbiased estimators of the population parameter λ.



CHECK YOUR PROGRESS

Q 1: State true or false.

(i) The estimating function of a parameter is the estimator. (T/F)

(ii) An estimate T_n is said to be an unbiased estimate of the parameter θ, if $E(T_n) > \theta$. (T/F)

(iii) In sample from a normal population, the sample mean is an unbiased estimate of the population mean θ. (T/F)

(iv) In a random sample from a normal population, the sample variance is an unbiased estimate of the population variance. (T/F)

Q 2: Distinguish between estimator and an estimate.

Q 3: What is the difference betwen point estimator and an interval estimator.

11.5.2 Consistency

The property of an estimator being consistent is derived from the law of large numbers which states that, as sample size increases, the sample mean $\bar{x}$ tends to the population parameter μ. similarly, a point estimator is consistent, if its value, say, $\bar{x}$ tend to become closer to the population parameter μ, as the sample size increases. So we can define a consistent estimator as follows :

Let $\hat{\theta}_n$ be an estimate of θ. based on a sample of size n, then $\hat{\theta}_n$ will be a consistent estimator of θ, if,

$$P\left\{\hat{\theta}_n \to \theta\right\} \to 1 \text{ as } n \to \infty$$

That is the probability that $\hat{\theta}$ approaches θ is 1 as n becomes larger and larger, i.e. higher the sample size, more concentrated should be the distribution be around θ.

For a symmetrical distribution, we know that both sample mean $\bar{x}$ and sample median x_{med} are unbiased estimators for a population mean μ. so, in case of a symmetrical distribution, consistency is the only desirable criterion to test for a good estimator. so, we see that the sample mean $\bar{x}$ is a consistent estimator of the population mean μ because $v(\bar{x}) = \frac{\sigma^2}{x}$ tends to zero for large n.

Again, the sample proportion $\bar{p}$ is a consistent estimator of the population proportion P since $v(\bar{p}) = \frac{\sigma^2}{n} \to 0$ as $n \to \infty$

As a matter of fact, it is seen, that consistency is the desirable property of an estimator when the sample size is large. It is of no pratical signifcance when the sample is small.

Remark :

(i) Consistency and unbiasedness of an estimator doesnot necessarily go together, i.e an unbiased estimator need not be consistent and a consistent estimator may be biased.

(ii) Consistent estimators are not unique.

1. If $x_1, x_2 ...., x_n$ is a random sample from a Normal population $N(\mu_1 1)$, show that $t = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2}$, is an unbiased estimator of $\mu^2 + 1$.

Solution: since $x_1, x_2 ...., x_n$ is the sample from $N(\mu_1 1)$

$$\therefore E(x_1) = \mu, \; V(x_1) = 1 \quad \forall i = 1,2,..n$$

Now, $V(X_i) = E(X_i^{\,2}) - \{E(x_i)\}^2$

$$\Rightarrow E(X_i^{\,2}) = V(X_i) + \{E(X_i)\}^2$$

$$= 1 + \mu^2$$

Now, E(t) $= E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2}\right)^2$

$$= \frac{1}{n}\sum_{i=1}^{n} E(x_i^{\,2})$$

$$= \frac{1}{n}\sum_{i=1}^{n} (1 + \mu^2)$$

$$= (1 + \mu^2)$$

2. Prove that in sampling from a $N(\mu, \delta^2)$ population, the sample mean is a consistent estimator of μ.

Solution: We know that, in sampling from $N(\mu, \delta^2)$ population, the sample mean $\bar{x}$ is also normally distributed with mean μ and variance $\frac{\delta^2}{n}$.

$$\bar{x} \sim N\left(\mu, \frac{\delta^2}{n}\right)$$

Now, $\frac{\lim}{n \to \alpha} V(\bar{x}) = \frac{\lim}{n \to \alpha} \frac{\delta^2}{n} = 0$

$\therefore \bar{x}$ is consistent estimator of μ.

11.5.3 Efficiency

The efficiency f an estimator is measured in terms of it's variance : for the same population, if there are two consistent

estimators for a parameter, then that consistent estimator with smaller variance is said to be more efficient, because it provides an estimate closer to the population parameter. Thus for a random sample of size n, if $\bar{\theta}_1$ and $\bar{\theta}_2$ are two consistent estimators of the population parameter θ, then the relative efficency of $\bar{\theta}_2$ compared to $\bar{\theta}_1$ is given by.

$$E = \frac{v(\bar{\theta}_1)}{v(\bar{\theta}_2)}$$

obviously, E < 1.

Where, $\bar{\theta}_1$ is the most efficient estimator.

By most efficient estimator, we mean that estimator in a class of consistent estimator for a parameter, whose variance is less than that of any such estimator.

Illustrated Example 1. In a normal sample from a $N(\mu, \sigma^2)$ population the sample mean $\bar{x}$ is the most efficient estimator for estimating.

μ and $V(\bar{x}) = \frac{\sigma^2}{n}$

Let x_{med} be the median for this sample.

then $V(x_{med}) = \frac{\pi\sigma^2}{2n}$

Efficency of $x_{med} = \frac{V(\bar{x})}{V(x_{med})} = \frac{\sigma^2/n}{\pi\sigma^2/2x} = \frac{2}{\pi}$

= 0.64

Again, whenever we deal with unbiased estimators of a parameter, we prefer that estimator which has minimum variance. This leads us to the minimum variance unbiased estimator of θ.

Remark : The correlation co-efficient between a most efficient estimator and any other estimator with efficiency e is $\sqrt{e}$.

Sufficicncy: An estimator is said to be sufficient for a parameter, if it contains all the information in the sample regarding the parameter. That is sufficiency required that an estimator uses the entire sample

informations in estimating the parameter of our interest.

For example the sample statistic $\bar{x}$ to be a sufficient estimator of the population parameter, say, μ, it must contain the maximum information and furnish as much of it as possible it obtaining the value of the parameter μ. A sufficient estimator, if available is always preferred over a non-sufficient estimator, because it allows the use of comprehensive sample information in estimation, so if we say that the sample mean $\bar{x}$ is sufficient estimator of μ, it implies that no other estimator, say x_{med} can convey more information in estimating the parameter μ.

Ex : In a normal population with mean μ and variance σ^2 we have,

(i) $\bar{x} = \frac{1}{n}\sum x_i$ is a sufficient estimator for μ, where σ^2 known,

(ii) $= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ is sufficient estimator for σ^2.



CHECK YOUR PROGRESS

Q 4: State true or false.

(i) In sampling from a normal population, the sample mean is a consistent estimator of the population mean μ. (T/F)

(ii) A consistent estimator is unique. (T/F)

(iii) A consistent estimator T of parameter θ is the most efficient estimator iff V(T) < V ($T^{/}$), where $T^{/}$ is any other consistent estimator of θ. (T/F)

(iv) If T_1 is the most efficient estimator of a parameter θ and T_2 be another estimator of θ, these the efficiency e of T_2 is, $e = \frac{V(T_2)}{V(T_1)}$

(T/F)

Q 5: Prove that, if $T^{//} = \frac{1}{2}(T + T^{/})$, where T and T/ are both efficient

estimators with variance V, then, $V(T'') = \frac{1}{2}V(1+p)$.

Where p is the correlation co-efficient between T and T/.

Q 6: Define consistent estimator. Do consistency and unbiasedness in estimators necessarily go together.



11.6 LET US SUM UP

- Estimation is the technique of coming to a decision about the value of the parameter depending on the imformation provided by the random sample.
- Estimation may be of two types - point estimation and interval estimation.
- The estimating function of a parameter is known as estimator.
- The particular value of an estimator is called the estimate of the parameter.
- The criteria for a good estimator are :
  unbiasedness
  consistency
  efficiency and sufficiancy
- An estimator T is said to be unbiased estimator of the parameter θ, if, E(T) = θ.
- θ_n is a consistent estimator of θ if θ_n converges to θ in probability for large n, where n is the sample size.
- An unbiased estimator need not be consistent and a consistent estimator may be biased.
- The sample mean is an unbiased estimator of the population mean.
- Sample variance is not an unbiased estimator of the population variance of
- If θ_1 is the most efficient estimator with variauce V_1 and θ_2 is any other consistent estimator with variauce V_2 then efficiency of θ_2 is.

$e = \frac{V_1}{V_2}$, e < 1.

- The correlation co-efficient between a most efficient estimator and any other estimator with efficiency e is $\sqrt{e}$.
- t is called minimum variance unbiased estimor of θ, if.

  (i) $E(t) = \theta$ (ii) $V(t) < V(t')$

  where t' is any other unbiased estimator of θ.



11.7 FURTHER READING

1) Gupta, S. C. and Kapoor, V. K (2002) Fundementals of Mathematical Statistics, Sulten Chand and sons.

2) Goon, A. M, Gupta, M. (1985). An Outline of Statistical Theory, Vol I, New World Press Pvt. Ltd.

3) Mukhopadhyay, P (1996), Mathematical Statistics, New Central Book Agency (P) Ltd.

4) Rao, C. R. (1963), Linear Statistical Inference and its Application, John Willey and sons.



11.8 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: (i) True, (ii) False, (iii) True, (iv) False.

Ans to Q No 2: The estimator is the estimated formula of a parameter. Again, the numerical value of an estimator from a fixed set of values of a random sample is known as estimate. Thus estimate stands for the value of the parameter. For example, the sample mean $\bar{x}$ is an is an estimate of population mean μ and sample mean square s^2 is an estimate of the population variance σ^2, whereas the formula for $\bar{x}$ and s^2 are the estimators.

Ans to Q No 3: The point estimation comprises of finding out an estimated formula of a parameter and finding out the numerical value on the basis of a sample using that estimated formula.

On the otherhand, interval estimation is the method of finding an interval which covers the true parameter with an assigned probability, say $100(1-\alpha)\%$. Such an interval is called a confidence interval with confidence co-efficient $(1-\alpha)$.

Ans to Q No 4: (i) True, (ii) False, (iii) True, (iv) False.

Ans to Q No 5: $V(T^{//}) = V\left[\frac{1}{2}(T+T^{/})\right]$

$$= \frac{1}{4}V(T+T^{/})$$

$$= \frac{1}{4}\left[V(T)+V(T^{/})+2\operatorname{cov}(T.T^{/})\right]$$

$$= \frac{1}{4}\left[V+V+2\rho V\right] \qquad \because V(T)=V(T^{/})=V$$

$$\rho = \frac{\operatorname{cov}\left(T,T^{/}\right)}{\sqrt{v(T)}\sqrt{v\left(T^{/}\right)}}$$

$$\therefore \quad \operatorname{cov}\left(T,T^{/}\right)=\rho V$$

$$\therefore \quad v\left(T^{//}\right)=\frac{1}{4}\left[2v(1+\rho)\right]$$

$$= \frac{1}{2}\left[v(1+f)\right]$$

Ans to Q No 6: Let θ_n be an estimator of θ based on a sample of size n, then θ_n is a consistent estimator of θ, if θ_n converges in probability to θ, for. large n, i.e.

$\theta_n \overset{p}{\to} \theta$ as $n \to \alpha$

Again, consistency and unbiasedness of an estimator doesnot necessarily go together. An unbiased estimator neednot necessarily be consistent. similarly, a consistent estimator may be biased.



11.9 MODEL QUESTIONS

Q 1: Define an estimator and an estimate. State the requirements of a good estimator.

Q 2: Define the following terms and give one example for each,

(i) Consistency,

(ii) Unbiasedness

(iii) Efficiency.

Q 3: Show that the unbiased estimators are not unique with the help of an example.

*** ***** ***

UNIT 12: INDEX NUMBERS

UNIT STRUCTURE

12.1 Learning Objectives

12.2 Introduction

12.3 Index Numbers: Concept and Uses

12.4 Construction of Simple Index Numbers

12.5 Construction of Weighted Index Numbers

- 12.5.1 Laspeyres' Index Number
- 12.5.2 Paasche's Index Number
- 12.5.3 Fisher's Index Number

12.6 Test of Adequacy of Index Numbers

- 12.6.1 Time Reversal Test
- 12.6.2 Factor Reversal Test
- 12.6.3 Circular Test

12.7 Base Shifting, Splicing and Deflating

12.8 Index of Industrial Production (Concept only)

12.9 Let Us Sum Up

12.10 Further Reading

12.11 Answers To Check Your Progress

12.12 Model Questions

12.1 LEARNING OBJECTIVES

After going through this unit, you will be able to:

- describe the uses of index numbers
- discuss the simple and weighted index numbers
- evaluate test of accuracy of index numbers
- discuss the concept of base shifitng, splicing and deflating.

12.2 INTRODUCTION

This unit deliberates on index numbers. Index numbers are the indicators of the various trends in an economy. Price index numbers indicate the position of prices whether they are rising or falling and at what rate.

Similarly, index numbers regarding agricultural production indicates the trend of change whether it is rising or falling at what rate over a period of time. An index number is an economic data figure reflecting price or quantity compared with a standard or base value. The base usually equals 100 and the index number is usually expressed as 100 times the ratio to the base value.

After going through this unit, you will be able to discuss the different concepts relating to index numbers, construct simple and wighted, discuss the concepts of base shifting, splicing and deflating of index numbers.

12.3 INDEX NUMBERS: CONCEPT AND USES

Dear learners, you might have often come across situations like the people making hue and cry over rising prices of goods and commodities. How do we analyse the rising prices in Economics? What tools do we use? To analyse the rising prices, rather change in the price levels, we use the tool 'index numbers' in Economics. Index numbers are one of the most frequently used statistical tools in the field of Economics and other social sciences. Index numbers represents the 'state of happening' in certain variables like the general price level, the wholesale price level, wage rate of labours, retail sales volume, wholesale sales volume, export and imports of a country etc. On such accounts, the index numbers are aptly termed as 'barometers of economic activity'.

We can define an index number as a statistical measure designed to show changes in a variable or a group of related variables with reference to time, geographic location or other entities like income, profession etc. As a statistical measure, an index number is calculated as a ratio of the current value to a base value (also called as reference value) and is expressed in percentage. It should be kept in mind that the index number is more commonly called as an index.

Thus, an index number is a statistical device (measure) with a purpose of showing average changes in one or more related variables (like price or quantity) between two periods of time (say, between 1991 and 1996) or two places like cities (say, Delhi and Mumbai) or countries (say, India and

Japan). For example, we may be interested in knowing which city of India out of Delhi, Mumbai, Kolkata and Chennai is the costliest or the cheapest in terms of price level. A tourist may be interested in knowing about the cost of living at different tourist places. Consumer price index number or cost of living index number helps in taking such decisions.

Let us consider the following table showing the changes in the wholesale prices of primary articles, manufactured products and commodities across all categories in India with reference to the price levels of 1981-82. This means, 1981-82 is the base year prices, i.e., (the index of prices of 1981-82=100).

Table 12.1: Index Numbers of Wholesale Prices in India

Base year 1981-82 (=100)

Period	Primary articles	Manufacture products	All commodities
July, 1992	237.6	222.6	226.6
October, 1992	237.1	229.3	232.4
January, 1993	235.0	228.6	231.6
April, 1993	234.2	231.2	234.6
June, 1993	234.9	233.5	237.7

Source: Centre for Monitoring Indian Economy, July, 1993.

From Table 12.1 it can be seen that compared to the 1981-82 price levels, the price level of primary goods (primary goods comprises food items, non-food items and minerals) in January, 1993 has increased by 234.2 percent or 2.342 times of the 1981-82 price levels of the corresponding period (i.e., January, 1982).

Thus, we have seen that index numbers give us a comparative picture over a period of time. This means that index numbers essentially reflects upon the changes in the variable or the group of related variables over a period of time. For example, if the index of industrial production was 215.1 in 1992-93 (base year: 1980-81), it means that the industrial production in that year increased by 2.151 times compared to 1980-81. It should however be noted that this does not mean that compared to 1980-81, all the sectors of the industry witnessed similar trend of progress. The index of 1992-93 reflects an 'average picture' of the industrial sector. It does not necessary imply that all the sectors of the economy witnessed similar

progress. This means that a few sectors of the industrial sectors might have performed above the average level, while a few might have performed below that level. According to Croxton and Cowden, "Index numbers are devices which measures the change in the level of a phenomenon with respect to time, geographical location or some other characteristics."

It should be kept in mind that different commodities are measured in different units. For example, rice and wheat are measured in kilograms, cloth in meters and milk in liters etc. Index numbers attempt some averages relating to commodities which are measured in different units. For purpose of comparison, we are interested in knowing relative changes. These relative changes are expressed in percentage terms.

Uses of Index Numbers: Index numbers are constructed with various aims and purposes. Index numbers constructed are put into several uses. Some of them are described below:

- The most commonly used types of index numbers are those that relate to changes in level in prices of a commodity or a group of commodities. Such index numbers are price index numbers. Price index numbers are useful in studying price movements with a view to analyzing their causes, as well as to determine their effects on the economy. To determine economic relationships, it is often useful to compare changes in the market price level with changes in other series, such as gold, bank deposits, bank loans, and the physical volume of production.
- Indices of physical changes over a period of time in production, in marketing, in sales, in imports and exports, and so on have important uses. Such index numbers over a period of time are known as index time series. These index numbers are extremely useful in the study of the movement of the characteristics over time, in the study of seasonal and cyclic behaviour, and in the study of business cycles, etc.
- Index numbers can be used for forecasting future trend in demand of commodities or in volume of production etc. Index numbers of industrial and agricultural production not only reflect the trend but also are useful in forecasting production. Index numbers of unemployment which

reflect the trend of unemployment are useful in determining factors leading to unemployment and in forecasting future trends.

- Index numbers which reflect the general price level, and in particular, the consumer price index numbers are useful in determining the quantum of additional wages or dearness allowances to compensate for the changes in the cost of living. The Government of India and most of the State Governments have index linked formulae for grant of dearness allowances to their employees.
- The changes in consumer price index can be used to determine the real income over a period. Suppose that a person's income in 1980 was 150 times than that in 1970, that is, there is an increase of 50% in money income. Suppose that the cost of living index number in 1980 was 175 times than that in 1970, that is, there is an increase of 75%. Then the person's real income in 1980 is $\frac{150}{175} = 85.7\%$ of his income in 1970. Thus index numbers are useful in determining the real income.



CHECK YOUR PROGRESS

Q 1: State whether the following statements are True (T) or False (F).

(i) Index numbers are specialized averages to measure relative changes. (T/F)

(ii) Index numbers measure net change in related variables over time. (T/F)

(iii) Simple index numbers require weights in their construction. (T/F)

(iv) Index numbers do not indicate the trend of change in the economy. (T/F)

Q 2: Fill in the blanks:

(i) Index number is a statistical device to express ..... change in related variable.

(ii) An index number is a ..... average.

12.4 CONSTRUCTION OF SIMPLE INDEX NUMBERS

In general we are interested in general price level rather than on the price level of any particular item. So we have to consider a number of items at the same time. If each commodity is of different qualities then all such qualities have also to be taken into consideration. We have to consider simultaneously a number of items or commodities together for measuring change in price or production.

The sources of data must be reliable and care should be taken in selecting the sources of data. For example, official publications as well as other reliable unofficial reports from producers and merchants should be considered for computing index for production. If retail food prices are being considered, then prices in stores, not in wholesale markets, are to be taken into consideration.

Now we shall discuss how to construct simple index numbers from a given set of data, like prices of certain commodities, production of certain commodities etc.

There are two methods of constructing simple index numbers. These are:

- Simple aggregate method and
- Method of simple average of relatives.

Remember that all items in a simple index number are assigned equal weights. These are unweighted index numbers.

- **Simple Aggregate Method:** In this case each item is given equal weights. If we give an equal weight to each item it means the same thing, whether each item is given a weight or not. It is the simplest method of constructing index numbers.

  Simple aggregate method to find price index numbers: We use the following three steps to find price index number.

  Step 1: Find the sum of the current year (that is, given period) prices of all the items included in the list. If 1 denote the current year, then find $\sum p_1$.

  Step 2: Find the sum of the base year prices of all the items included in

the list. If 0 denote the base year, then find $\sum p_0$.

Step 3: Use the formula $p_{0/1} = \frac{\sum p_1}{\sum p_0} \times 100.$

Exercise 1: Find the price index number, using simple aggregate method, for the following information.

Commodities	Units	Prices (in Rupees)	
		2000 (year)	2005 (year)
Wheat	Quintal	800	1200
Rice	Quintal	1200	1600
Milk	Litres	12	18
Tea	Kilograms	100	120
Clothing	Metres	30	40
Meat	Kilograms	80	120

Solution: Represent prices of 2005 (current year) as p_1 and that of 2000 (base year) as p_0 and reconstruct the table as follows:

Commodities	p_0	p_1
Wheat	800	1200
Rice	1200	1600
Milk	12	18
Tea	100	120
Clothing	30	40
Meat	80	120
	$\sum p_0 = 2222$	$\sum p_1 = 3098$

Therefore, the price index of 2005 with base year 2000 is:

$$p_{0/1} = \frac{3098}{2222} \times 100 = 139.4$$

Thus the increase in prices (or price level) is 139.40 - 100 = 39.4%.

Exercise 2: The prices of three commodities A, B and C increased from Rs.50, Rs.7 and Rs.3 in 2005 to Rs.52, Rs.8 and Rs.5 respectively in 2006. Using the simple aggregate method, find by how much on an average the prices have increased?

Solution: We know construct the following table:

Commodities	Price in 2005 (p_0)	Price in 2006 (p_1)
A	50	52
B	7	8
C	3	5
	$\sum p_0 = 60$	$\sum p_1 = 65$

Therefore, the price index of 2006 with respect to the base period 2005 is:

$$p_{2005/\ 2006} = \frac{\sum p_1}{\sum p_0} \times 100 = \frac{65}{60} \times 100 = 108.33$$

This shows that the price index of current year 2006 with respect to base year 2005 is 108.33%. Hence the price increase in one year is 108.33 - 100 = 8.33%.

Simple aggregate method to find quantity index numbers: We use the following three steps to find quantity index number.

Step 1: Find the sum of the current year (that is, given period) quantity or volume produced, consumed or transacted of all the items included in the list. If 1 denote the current year, then find $\sum q_1$.

Step 2: Find the sum of the base year quantity or volume produced, consumed or transacted of all the items included in the list. If 0 denote the base year, then find $\sum q_0$.

Step 3: Use the formula $q_{0\ /\ 1} = \frac{\sum q_1}{\sum q_0} \times 100.$

- **Method of Simple Average of Relatives:** In this method, average of the relatives is obtained by using any one of the measures of central tendency. In particular, here we shall use arithmetic mean for averaging the relatives. We shall discuss briefly this method only for price relatives. Similar formulas on quantity index number and value index number can be obtained considering quantity and value relatives instead of price relative. There are few questions to find quantity and value index numbers in the section "Possible Questions". You should work out those problems with the help of the problems on price index numbers which are solved below.

Method of Simple Average of Price Relatives:

We know that a price relative is nothing but the ratio of current year prices to those in the base year. Taking the arithmetic mean of the price relatives, we get the following formula for price index number:

$$p_{0/1} = \frac{\sum\left(\frac{p_1}{p_0} \times 100\right)}{N}$$

where N stands for number of commodities included in the index numbers. Steps to calculate price index numbers by Simple Average of Price Relatives Method:

Step 1: Find percentage price relative for each commodity, that is, $\frac{p_1}{p_0} \times 100$

Step 2: Find the sum of these percentage price relatives, that is, $\sum\left(\frac{p_1}{p_0} \times 100\right)$.

Step 3: Divide $\sum\left(\frac{p_1}{p_0} \times 100\right)$ by the total number of commodities included in the list.

Exercise 3: Give the solution of Exercise 1 using the Simple Average of Price Relatives.

Solution: We construct the following table:

Sl No.	Commo-dities	Units	Prices (in Rupees)		Price Relative $\frac{p_1}{p_0} \times 100$
			2000 p_0	2005 p_1	
1	Wheat	Quintal	800	1200	$\frac{1200}{800} \times 100 = 150.00$
2	Rice	Quintal	1200	1600	$\frac{1600}{1200} \times 100 = 133.33$
3	Milk	Litres	12	18	$\frac{18}{12} \times 100 = 150.00$
4	Tea	Kilograms	100	120	$\frac{120}{100} \times 100 = 120.00$

5	Clothing	Metres	30	40	$\frac{40}{30}\times 100 = 133.33$
6	Meat	Kilograms	80	120	$\frac{120}{80}\times 100 = 150.00$
N = 6					$\sum\left(\frac{p_1}{p_0}\times 100\right) = 836.66$

Hence, the price index number of 2005 with respect to base year 2000 is:

$$p_{0/1} = \frac{\sum\left(\frac{p_1}{p_0}\times 100\right)}{N} = \frac{836.66}{6} = 139.44.$$

Exercise 4: Give the solution of Exercise 2 using the Simple Average of Price Relatives.

Solution: We first prepare the following table:

Sl No.	Commo-dities	Prices (in Rupees)		Price Relative
		2005 p_0	2006 p_1	$\frac{p_1}{p_0}\times 100$
1	A	50	52	$\frac{52}{50}\times 100 = 104.00$
2	B	7	8	$\frac{8}{7}\times 100 = 114.29$
3	C	3	5	$\frac{5}{3}\times 100 = 166.66$
N=3				$\sum\left(\frac{p_1}{p_0}\times 100\right) = 384.95$

Hence, the price index number of 2006 with respect to base year 2005 is:

$$p_{0/1} = \frac{\sum\left(\frac{p_1}{p_0}\times 100\right)}{N} = \frac{384.95}{3} = 128.32.$$

12.5 CONSTRUCTION OF WEIGHTED INDEX NUMBERS

In weighted index number each item is given weight according to the importance it occupies in the list. There are two groups of methods to calculate index number of this category: (a) Weighted aggregate method and (b) Weighted average of price relative methods. In this unit, we will study three types of weighted aggregate method. These methods are popularly known as (1) Laspeyre's Method, (2) Paasche's Method and (3) Fisher's Method. All the methods are used to calculate weighted index numbers. The main difference between the Laspeyre's Method and Paasche's Method is that Laspeyre's uses base year quantities of commodities as their relative weights, while Paasche's uses current year quantities of commodities as their relative weights for preparing a price index.

12.5.1 Laspeyre's Index Number

Laspeyre's Price Index Number: It uses base year quantities as the weights. Accordingly, the formula is

$$P_L = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

Laspeyre's Quantity Index Number: If we multiply the quantities by the base year price, then we get Laspeyre's quantity index number. Accordingly the formula for Laspeyre's quantity index number is:

$$Q_L = \frac{\sum q_1 p_0}{\sum q_0 p_0} \times 100$$

Steps to Calculate Weighted Index Number by Laspeyre's Method:

Step 1: Multiply current year price (p_1) with the base year quantity (q_0) to get $p_1 q_0$ for each item/ commodity.

Step 2: Multiply base year price (p_0) with the base year quantity to get $p_0 q_0$ for each item/ commodity.

Step 3: Add all $p_1 q_0$ and $p_0 q_0$ separately to get $\sum p_1 q_0$ and $\sum p_0 q_0$.

Step 4: Divide $\sum p_1 q_0$ by $\sum p_0 q_0$ and multiply by 100 to obtain Laspeyre's price index number.

Exercise 5: Find Laspeyre's price index number from the following data.

Commodity	Price (in Rs)		Quantity bought in units	
	1995	1996	1995	1996
A	50	52	10	12
B	7	9	3	4
C	3	5	7	6

Solution: Represent prices and quantities in current year by p_1 and q_1 respectively and represent prices and quantities in base year by p_0 and q_0 respectively. Let us construct the following table following the steps 1-3.

Commodities	p_0	q_0	p_1	p_0q_0	p_1q_0
A	50	10	52	500	520
B	7	3	9	21	27
C	3	7	5	21	35
				$\sum p_0q_0$ = 542	$\sum p_1q_0$ = 582

Therefore the Laspeyre's price index number is:

$$P_L = \frac{582}{542} \times 100 = 107.38$$

12.5.2 Paasche's Index Number

Paasche's Price Index Number uses current year quantities *(q_1)* as the weights. Accordingly, the formula is

$$P_P = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

Paasche's Quantity Index Number: It uses current year prices *(p_1)* as the weights. Accordingly, the formula is

$$Q_P = \frac{\sum q_1 p_1}{\sum q_0 p_1} \times 100$$

Exercise 6: Find Paasche's price index number from the data given in Exercise 5.

Solution: Represent prices and quantities in current year by p_1 and q_1 respectively and represent prices and quantities in base year by p_0 and q_0 respectively. Let us construct the following table.

Commodities	p_0	q_1	p_1	p_0q_1	p_1q_1
A	50	10	52	500	520
B	7	3	9	21	27
C	3	7	5	21	35
				$\sum p_0q_1$ = 646	$\sum p_1q_1$ = 690

Therefore the Paasche's price index number is:

$$P_P = \frac{690}{646} \times 100 = 106.8$$

Remark: You have noticed that the value of Laspeyre's index number (107.38) and the value of Paasche's index number (106.8) are different for the same data. The difference has arisen because of differences in weights.

12.5.3 Fisher's Index Number

The Fisher's price index number is nothing but the geometric mean of the Laspeyre's and Paasche's price index numbers. Hence the formula for Fisher's price index number is:

$$P_F = \sqrt{P_L \times P_P} = \sqrt{\left(\frac{\sum p_1q_0}{\sum p_0q_0} \times 100\right) \times \left(\frac{\sum p_1q_1}{\sum p_0q_1} \times 100\right)}$$

$$= \sqrt{\frac{\sum p_1q_0}{\sum p_0q_0} \times \frac{\sum p_1q_1}{\sum p_0q_1}} \times 100 \qquad \cdots(1)$$

Similarly, the Fisher's quantity index number is given by:

$$Q_F = \sqrt{Q_L \times Q_P} = \sqrt{\left(\frac{\sum q_1p_0}{\sum q_0p_0} \times 100\right) \times \left(\frac{\sum q_1p_1}{\sum q_0p_1} \times 100\right)}$$

$$= \sqrt{\frac{\sum q_1p_0}{\sum q_0p_0} \times \frac{\sum q_1p_1}{\sum q_0p_1}} \times 100 \qquad \cdots(2)$$

Exercise 7: Find Fisher's price index number from the data given in Exercise 5.

Solution: We have already found that for the given data,

$P_L = 107.38$ and $P_P = 106.8$.

Hence the Fisher's price index number is:

$$P_F = \sqrt{107.38 \times 106.8} = 107.09$$



CHECK YOUR PROGRESS

Q 3: State whether the following statements are true (T) or False (F).

(i) There are two types of index numbers: simple or unweighted and weighted index numbers.(T / F)

(ii) There are two methods of constructing simple index numbers: simple aggregate method and simple average relatives. (T / F)

(iii) We the same index number whether we use simple aggregate method or simple average of price relatives' method. (T / F)

(iv) Laspeyre's uses current year quantities as weights. (T / F)

(v) Paasche's uses base year quantities as weights. (T / F)

Q 4: Fill up the blanks.

(i) In Laspeyre's method of finding price index number we use ..... quantities as weights.

(ii) In Paasche's method of finding price index number we use ..... quantities as weights.

(iii) $I = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$ is ..... weighted index number.

(iv) $I = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$ is ..... weighted index number.

12.6 TEST OF ADEQUACY OF INDEX NUMBERS

The index numbers as an aggregative relative should satisfy a few properties. We shall examine the properties that a good index number should have.

12.6.1 Time Reversal Test

If the two periods, the base period and the reference period are interchanged, the product of the two index numbers should be unity. In other words, one should be the reciprocal of the other.

Denote the index number of the given year (1) relative to the base year (0) by $I_{0/1}$. Then $I_{1/0}$ is the index number of the year (0) relative to the year (1), that is, with the years interchanged time reversal test requires that

$$I_{0/1} \times I_{1/0} = 1, \qquad or \qquad I_{1/0} = \frac{1}{I_{0/1}}$$

An index number which satisfies the above property is said to satisfy time reversal test.

Remark: As we have seen in section 19.5, all the relatives satisfy the time reversal property.

Let us now examine whether Laspeyre's index number satisfy this property or not. We know that Laspeyre's index number (without percentage) is $\frac{\sum p_1 q_0}{\sum p_0 q_0}$. Interchanging the time subscripts, we get $\frac{\sum p_0 q_1}{\sum p_1 q_1}$ and hence their product is $\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_0 q_1}{\sum p_1 q_1} \neq 1$.

Consequently, Laspeyre's price index number does not satisfy the time reversal property. Similarly, it can be seen that Laspeyre's quantity index number also does not satisfy the time reversal property. Also, Paasche's price and quantity index numbers do not satisfy this property.

Now consider the Fisher's price index number. Interchanging the time subscripts we find another index number. The product of the two is given by

$$\sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times \sqrt{\frac{\sum p_0 q_1}{\sum p_1 q_1} \times \frac{\sum p_0 q_0}{p_1 q_0}}$$

which is equal to 1. Hence Fisher's price index number satisfies the time reversal property.

12.6.2 Factor Reversal Test

If the two factors *p* and *q* in a price index formula are interchanged, so that a quantity index number is obtained, then the product of the two index numbers should give the true value ratio

$$\frac{\sum p_1 q_1}{\sum p_0 q_0}$$

In other words, the price index number multiplied by the corresponding quantity index number should give the true ratio of value in the given year (1) to the value in the base year (0). This property holds for a single commodity, since

$$\frac{p_1}{p_0} \times \frac{q_1}{q_0} = \frac{p_1 q_1}{p_0 q_0}$$

but it does not hold for most of the index numbers. For example,

$$P_L \times Q_L = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum q_1 p_0}{\sum q_0 p_0} \neq \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

Now consider the Fisher's index number. We have from the formulas (1) and (2),

$$P_F \times Q_F = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \times \frac{\sum q_1 p_1}{\sum q_0 p_1}} = \frac{\sum p_1 q_1}{\sum p_0 q_0}$$

This proves that Fisher's index number also satisfy factor reversal test. Fisher's index number is one of the rare index numbers which satisfy both the time reversal and factor reversal tests. So it is called "Fisher's Ideal Index Number".

12.6.3 Circular Test

This test is based on the shifting of the base period. If $I_{i/j}$ denotes the index number for the given period *j* with respect to the base period *i* then this test requires that

$$I_{a/b} \times I_{b/c} \times I_{c/a} = 1$$

Remark: As we have seen in section 19.5, all the relatives satisfy the circular property.

Remark: Verify that none of the weighted index numbers satisfy circular test.

12.7 BASE SHIFTING, SPLICING AND DEFLATING

In this section, we shall briefly discuss the concepts of base shifting, splicing and deflating.

Base Shifting: In Statistics, we often require to change the reference base of an index number series from one time period to another without returning to the original raw data and recomputing the eniture series. The change in reference base period is usually referred to as '*base shifting*'.



ACTIVITY 12.1

The Ministry of Statistics, Government of India has changed the base for the calculation of GDP from time to time. Undertake an exercise on tracking such changes since 1990.

Base shifting is undertaken for two important reasons. a) The previous base (i.e., the base year currently in practice) has become too old is not much relevant for making comparision in the present times. By shifting the series, the series is presented to be 'current' and more relevant to the present times. b) When different data are presented for comparision pertaining to the different time periods, it becomes necessary to first make them comparable with the 'time' dimension. For example, it would be

irrevalent to compare the GDP estimates of a country of 1850, 1920, 1950 and 2010 without any statistical treatment. For making the comparision effective, the data must be first worked out to one base year (say, 2010). This is particularly important when we compare them in one single graph.

Splicing: Splicing means combining two or more series of overlapping index numbers to obtain a single index number on a common base. Sometimes, a series of index numbers is available for a period of time, which has been substantially revised including shifting of the base year. It may also happen that the weights assigned to the series have become out dated. As such, new weights may be assigned and the series is worked out anew. In such cases, two series would result, the old one and the new one. Now, it may become necessary to convert the two series into one continuous series. This is done with the help of splicing.

Splicing can be done only if the index numbers are constructed with the same items and have an overlapping year. Splicing like base shifting will give accurate result of the index numbers, use simple geometric mean and satisfy the circular test. In splicing of index numbers, we find a common factor by which the spliced index number series is multiplied to get a common base.

Deflating: Deflating means adjusting, correcting or reducing a value which is inflated. Thus, in defilating, we make adjustments for the effect of changing price levels. You know that a rise in price levels essentially reduces the real purchasing power of money. So, to compare the value of money (or its purchasing power), in different time peirods, we need to adjust the rate of inflation during the given peirod. This is done through deflating. Deflating is a technique of converting a series of values calculated at current prices into constant prices of a given year. This is particularly desirable in the case of an economy which has inflationary trends because in such an economy, the increase in the prices of commodities over a period of years means a fall in their real incomes. For example, a rise in money income may not amount to a rise in real income. As such, it becomes necessary to adjust or correct nominal wages in accordance with the rise in the corresponding price index to arrive at the real income. The purchasing power is given by

the reciprocal of the index number and consequently the real income is obtained on dividing the money or nominal income by the corresponding appropriate price index and multiplying the result by 100.

12.8 INDEX OF INDUSTRIAL PRODUCTION (CONCEPT ONLY)

The Index of Industrial Production (IIP) is an index which shows the growth rates in different industry groups of the economy in a stipulated period of time. The IIP index is computed and published by the Central Statistical Organisation (CSO) on a monthly basis.

Description: IIP is a composite indicator that measures the growth rate of industry groups classified under,

(1) Broad sectors, namely, Mining, Manufacturing and Electricity

(2) Use-based sectors, namely Basic Goods, Capital Goods and Intermediate Goods.

Currently IIP figures are calculated considering 2004-05 as base year.



12.9 LET US SUM UP

- Index number is a statistical measure or device with a purpose of showing average change in one or more related variables over time and space.
- Price index numbers are more commonly used. It measures relative changes in prices over a time period.
- We can have either simple or weighted index number. Simple index is also called unweighted index number or index number with equal weights. In this unit, we have discussed about the simple index numbers as well as weighted index number.
- We have learnt about two methods of constructing simple index numbers: Simple aggregate method and Method of simple average of relatives. The formula to find index number using simple aggregate

method is given by $p_{0/1} = \frac{\sum p_1}{\sum p_0} \times 100.$

Again, the formula to find price index number using the method of simple averages of relatives is given by $p_{0/1} = \frac{\sum\left(\frac{p_1}{p_0} \times 100\right)}{N}$.

- We have also leant about weighted index numbers. There are two groups of methods to calculate index number of this category: (a) Weighted aggregate method and (b) Weighted average of price relative methods. We briefly discussed about three types of weighted aggregate method. These methods are popularly known as (1) Laspeyre's Method, (2) Paasche's Method and (3) Fisher's Method.
- We have also learnt about the test of adequacy of index numbers. We mentioned three properties that a good index number should have. Fisher's index number satisfies both the time reversal and factor reversal tests and hence it is known as Fisher's Ideal Index Number.



12.9 FURTHER READING

1. J. Medhi, *Statistical Methods, An Introductory Text*, New Age International (P) Limited, Publishers, 2nd edition, 2000.



12.10 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: (i) True (ii) True (iii) False (iv) False

Ans to Q No 2: (i) average (ii) specialized

Ans to Q No 3: (i) True (ii) True (iii) False (iv) False (v) False

Ans to Q No 4: (i) base year (ii) current year (iii) Laspeyre's



12.11 MODEL QUESTIONS

Q 1: Explain the meaning of an index number and its main characteristics.

Q 2: "Index numbers are economic barometers". Why are index numbers so called? Explain.

Q 3: Explain the method of construction of unweighted index numbers by Simple Aggregate Method.

Q 4: Explain the method of construction of unweighted index numbers by Simple Aggregate of Price Relative Method.

Q 5: The exchange rate of a US dollar was Rs.44.00 in July 2008 and was Rs.50.00 in February 2009. Find the price relative.

Q 6: The production of rice in Assam in the month of January 2008 was 3608 tonnes and that in February 2009 was 3700 tonnes. Find the quantity relative.

Q 7: The following table shows the prices of the commodities consumed in 1970 and 1978.

Commodities	Unit	1970 Price (Rs)	1978 Price (Rs)
A	kilogram	4.00	4.60
B	litre	8.00	8.75
C	metre	7.00	8.00
D	kilogram	14.00	16.00
E	litre	3.00	4.00

Find the price index number using:

(a) simple aggregate method

(b) method of simple aggregate of price relatives.

Q 8: The production of rice, milk, cloth, wheat and petrol are given in the following table. Taking 1980 as base year, compute quantity index number using:

(a) simple aggregate method

(b) method of simple aggregate of quantity relatives.

Commodities	Unit	Quantity produced	
		1980	1990
Rice	kilogram	400000	700000
Milk	litre	800000	905000
Cloth	metre	100000	105000
Wheat	kilogram	140000	160000
Petrol	litre	300500	400000

Q 9: What do you mean by tests of index numbers?

Q 10: Explain the important criterion or test that a good index number should satisfy.

Q 11: The prices and quantity of five commodities are given in the following table. Taking 2000 as base year, compute value index number using:

(a) simple aggregate method

(b) method of simple aggregate of value relatives.

Comm-odities	2000		2002	
	Price (Rs)	Quantity (Kg)	Price (Rs)	Quantity (Kg)
A	400	10	700	12
B	800	15	1905	17
C	100	5	125	5
D	140	7	160	7
E	300	6	400	6

Q 12: Find Laspeyre's and Paasche's price and quantity index number from the following data. Also find the Fisher's price and quantity index numbers.

Commo-dities	Base year		Current year	
	Price (Rs) p_0	Quantity (kg) q_0	Price (Rs) p_1	Quantity (kg) q_1
A	5	25	6	30
B	10	5	15	4
C	3	40	2	50
D	6	30	8	35

Q 13: The following table shows the price quantity of the commodities consumed in 2000 and 2006.

Commo-dities	2000		2006	
	Price (Rs)	Quantity (kg)	Price (Rs)	Quantity (kg)
I	400	35	460	35
II	800	10	1000	10
III	300	15	50	12
IV	4000	10	5000	8
V	50	20	300	20

Find the following index numbers.

(1) Laspeyre's Price and Quantity index numbers.

(2) Paasche's Price and Quantity index numbers.

(3) Fisher's Price and Quantity index numbers.

*** ***** ***

UNIT 13: TESTING OF HYPOTHESIS

UNIT STRUCTURE

13.1 Learning Objectives
13.2 Introduction
13.3 Statistical Hypothesis: Null Hypothesis and Alternative Hypothesis
13.4 Errors in Hypothesis Testing, Level of Significance and Critical Region
13.5 One -Tailed and Two-Tailed Tests
13.6 General Procedure for Hypothesis Testing
13.7 Testing of Hypothesis in case of Large Samples
13.7.1 Hypothesis Testing for Single Population Mean
13.7.2 Hypothesis Testing for Single Population Proportion
13.7.3 Hypothesis Testing for Difference between Two Population Means
13.7.4 Hypothesis Testing for Difference between Two Population Means
13.8 Let Us Sum Up
13.9 Further Reading
13.10 Answers To Check Your Progress
13.11 Model Questions

13.1 LEARNING OBJECTIVES

After going through this unit you will be able to

- establish null and alternative hypotheses about a population parameter
- understand the two types of errors; interpret the confidence level and the significance level
- develop hypothesis testing procedures for accepting or rejecting a null hypothesis
- discuss how and when to use the normal and t distributions for hypothesis testing

13.2 INTRODUCTION

In today's dynamic business environment, decision makers need might answers to certain questions in order to take optimum decisions. For example a pharmaceutical concern might be interested in assessing whether a new drug is really effective in curing a disease; a personnel manager might be interested to study the job satisfaction level of employees, a marketing manager might be interested in knowing the customer satisfaction about a particular product, etc. In all the situations, the concerned manager has to take wise decisions since such decisions involve the risk of taking wrong decisions. It is here that modern theory of probability plays a vital role in decision making and the branch of Statistics that develops the criterion for such decisions is known as testing of hypothesis. Hypothesis testing begins with an assumption commonly termed as "hypothesis" that a researcher makes about an unknown population parameter. Hypothesis testing is a well-defined procedure which enables us to decide objectively whether to accept or reject the hypothesis based on the information obtained from the sample.

13.3 STATISTICAL HYPOTHESIS: NULL HYPOTHESIS AND ALTERNATIVE HYPOTHESIS

A statistical hypothesis is some assumption or statement, which may or may not be true, about an unknown population parameter. For example, suppose the Vice President (HR) of a company wants to examine the effectiveness of a training programme which the company has organized for all its employees located at different places of the country. The Vice President (HR) will then make an assumption that "training has not enhanced efficiency". To test such claim statistically, sample data are collected and analyzed. The assumption is accepted or rejected on the basis of the sample findings.

There are two types of statistical hypothesis: (i) Null hypothesis and (ii) Alternative hypothesis. In hypothesis testing, to reach decisions we must start with a hypothesis, called null hypothesis which is symbolized H_0. The

null hypothesis asserts that there exists no (significant) difference between the sample statistic and the population parameter and whatever observed difference is there is merely due to fluctuations of sampling from the same population. In the words of R.A. Fisher, "Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true." The null hypothesis that makes a claim regarding the specific value of the parameter is generally expressed by any one of the following mathematical statements

(i) $H_0 : \mu = \mu_0$ (ii) $H_0 : \mu \leq \mu_0$ (iii) $H_0 : \mu \geq \mu_0$

where μ is the population mean and μ_0 denotes a hypothesized value of μ.

An alternative hypothesis, generally denoted by H_1, is the opposite of the null hypothesis. Thus any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis. It is obvious that both null and alternative hypotheses cannot be true simultaneously at the same time. In other words, when null hypothesis is found to be true, the alternative hypothesis is to be rejected or when null hypothesis is rejected, the alternative hypothesis must be true. For example, if we are interested in studying the hypothesis that the population has a specified mean, μ_0 (say), then we have the null hypothesis $H_0 : \mu = \mu_0$ and against this null hypothesis we may have any one of the following alternative hypotheses (depending upon the situation) given by,

(i) $H_1 : \mu \neq \mu_0 \quad (i.e., \mu > \mu_0 \text{ or } \mu < \mu_0)$

(ii) $H_1 : \mu > \mu_0$ (iii) $H_1 : \mu < \mu_0$

The alternative hypothesis given by (i) is called two-tailed alternative and the alternatives given by (ii) and (iii) are called right-tailed and left-tailed alternatives respectively. Accordingly, the nature of the alternative hypothesis enables us to decide whether we have to use an one-tailed (right or left) or two-tailed test.

Simple and Composite hypothesis: If the hypothesis completely specifies the population, then it is called simple hypothesis, otherwise it is called composite hypothesis. For example, from a normal population $N(\mu, \sigma^2)$, the hypothesis of the form

$H: \mu = \mu_0$ and $\sigma^2 = \sigma_0^2$ (both μ and σ^2 are specified)

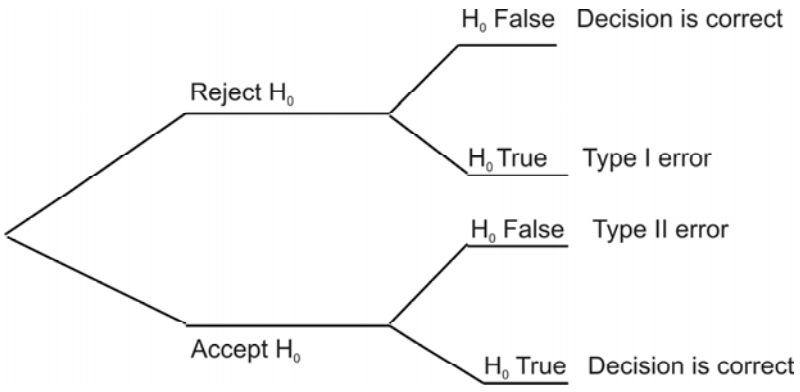
is a simple hypothesis because it completely specifies the distribution. On the other hand, the hypotheses of the forms

(i) $H: \mu = \mu_0$ (σ^2 is not specified), (ii) $H: \sigma^2 = \sigma_0^2$ (μ not being specified) are composite since each of these does not completely specify the value of the parameter.

13.4 ERRORS IN HYPOTHESIS TESTING, LEVEL OF SIGNIFICANCE AND CRITICAL REGION

In hypothesis testing, the null hypothesis H_0 should be accepted if it happens to be true and the null hypothesis H_0 should be rejected if it happens to be false. However, in practical situations, the correct decision is not always possible while testing a null hypothesis. Since the decision to reject or accept a null hypothesis depends only on sample data therefore a decision maker may commit two types of errors (i.e., take wrong decisions) which are diagrammatically shown in Fig.13.1.

Figure 13.1: Type I and Type II errors



Type I error is committed if we reject a null hypothesis when it is true. The possibility of committing Type I error is denoted by the symbol α. Thus,

$\alpha = P$ (Type I error)= P (Reject H_0 when H_0 is true)

On the other hand, Type II error is committed if we accept a null hypothesis when it is false. The possibility of committing Type II error is denoted by the symbol β. Thus we have

$\beta = P$ (Type II error)=(Accept H_0 when H_0 is false)

where α and β are also known as the sizes of Type I and Type II error respectively.

For example, a business firm, before introducing a new product, has to choose one of the two decisions- to introduce or not to introduce the product. Again if the product is introduced then the firm may face two situations- the failure of the product and the success of the product. The firm thus runs the risk of taking wrong decisions in two ways. The one risk of taking wrong decision known as Type II error, is that the firm does not introduce the product though it would have succeeded had it been introduced. The another risk of taking wrong decision known as Type I error is that the firm introduces the product but it failed.

In the terminology of Statistical Quality Control, Type I error is known as producer's risk while Type II error is known as consumer's risk.

In practice, it is not possible to make error free judgment in Statistics. Although both the errors should be very small yet it is not possible to minimize both these errors simultaneously at the same time on the same hypothesis test. Generally there exists an inverse relationship between α and β as a result of which an attempt to reduce type I error (α) leads to an increase in Type II error (β) and vice-versa. Since in most of decision making problems, it is more risky to accept a wrong hypothesis than to reject a correct one, i.e., consequences of Type II error are likely to be more serious than the consequences of Type I error therefore a compromise is made by minimizing the more serious error (i.e., β) after keeping fixed the less serious error (i.e., α).

In testing of hypothesis, the **Level of significance** is defined in terms of the probability of rejecting a null hypothesis when it is true. The level of significance is also known as the size of the critical region or the rejection region and generally denoted by α . It is to be noted that the level of significance should be determined before a test procedure so that obtained result is free from the influence of the decision maker. In practice, generally adopted level of significance by researchers is: 1 % (=0.01), 5% (=0.05) and 10% (=0.10). If a researcher use 5% level of significance, it would

mean that in 5 samples out of 100 he is likely to reject a correct null hypothesis H_0. In other words the researcher is 95% confident that his decision to reject H_0 is correct. Thus higher the significance level adopted for testing a hypothesis, the higher the chance of rejecting a true null hypothesis.

In hypothesis testing, the area under the sampling distribution curve of the test statistic is divided into two mutually exclusive regions termed as acceptance region and rejection or critical region. The acceptance region is a range of values of the sample statistic that lead us to the acceptance of null hypothesis H_0 while critical region is the range of sample statistic values within which if values of the sample statistic falls lead to the rejection of null hypothesis H_0. In Figure 13.2 if the calculated value of the sample statistic falls in the acceptance region, the null hypothesis is accepted otherwise it is rejected. For making a decision whether to accept or reject a null hypothesis, a researcher has to determine the critical value that separates the regions of acceptance and rejection. The determination of critical value depends on the size of the rejection region, which is directly associated with the risk involved in decision making.

13.5 ONE TAILED AND TWO TAILED TESTS

In any test, there are two types of tests of hypothesis. They are: (i) One-tailed tests and (ii) Two-tailed tests. Whether to use a one-tailed test or two tailed test depends entirely on the formulation of the alternative hypothesis. If the alternative hypothesis is one-tailed then the test to be applied is one-tailed and if the alternative hypothesis is two-tailed then the corresponding test is two-tailed. In case of two-tailed test, the critical or significant values of the test statistic lie towards both the tails of the graph of the sampling distribution of the statistic. Figure 13.2 relates to standard normal test statistic Z. The figure reveals that critical values of the standard normal test statistic Z. lies on both sides of the mean i.e., in both the tails of the distribution. The shaded region under the normal curve is the rejection region (or critical region) at 5% level of significance, the region being 5% of the total area. In other words, the total area under the normal curve being unity, the size of the rejection region is 0.05.

Again sometimes we may encounter certain situations where alternative hypothesis states that either one process is better than the other or one process is worse than the other. But in two-tailed test the alternative hypothesis H_1 states that both may occur i.e., one process may be better or worse than the other. Figure 13.3 and Figure 13.4 below exhibits the critical values and critical regions of the standard normal test statistic for right-tailed test and left-tailed test respectively for 5% level of significance.

Figure 13.2: Two-tailed Test

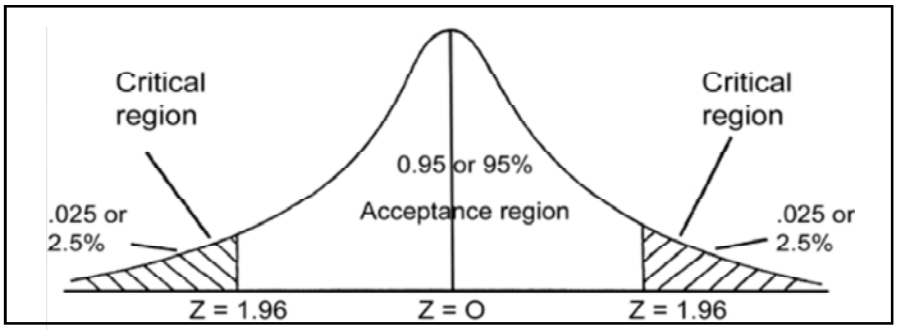


Figure 13.3: Right-tailed Test

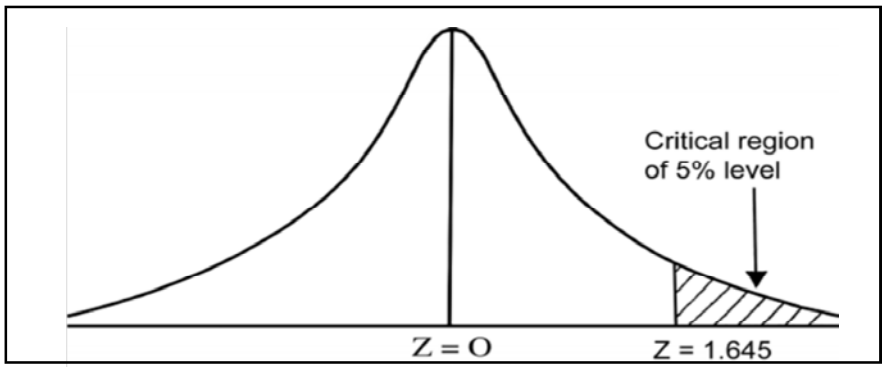
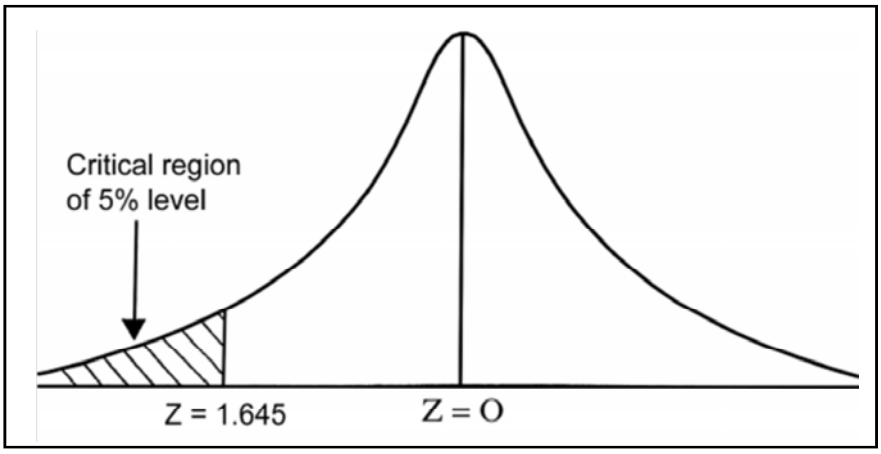


Figure 13.4: Left-tailed test



Now, we present below a summary of certain critical values of test statistic Z (i.e., standard normal test statistic) for both one-tailed and two-tailed tests at some commonly used level of significance. These critical values of Z for different levels of significance have been obtained by using the table of area under normal curve.

Table 13.1: Certain Critical Values of Test Statistic Z

Critical values (Z_α)	Level of significance (α)		
	5% or 0.05	1% or 0.01	10% or 0.010
Two-tailed test	$\lvert Z_\alpha \rvert = 1.96$	$\lvert Z_\alpha \rvert = 2.58$	$\lvert Z_\alpha \rvert = 1.645$
Right-tailed test	$Z_\alpha = 1.645$	$Z_\alpha = 2.33$	$Z_\alpha = 1.28$
Left-tailed test	$Z_\alpha = -1.645$	$Z_\alpha = -2.33$	$Z_\alpha = -1.28$

Note: (i) The notation Z_α stands for the critical value of Z at α level of significance.

(ii) Critical or significant values are also known as tabulated values.

(iii) If the size of the sample is small (usually less than 30) then the normality assumption of the sampling distribution of the test statistic is not valid and hence we cannot use the significant values given in the above table. In such situation we use the significant values based on the exact sampling distribution of Z, which turns out to be t, F or χ^2. (Please refer to any Standard textbook on Statistics).

13.6 GENERAL PROCEDURE FOR HYPOTHESIS TESTING

Hypothesis testing is a well-defined systematic procedure that involves five different steps as given below:

(1) The first step is to set up the null hypothesis H_0.

(2) Set up the alternative hypothesis H_1. The nature of alternative hypothesis H_1 helps us to choose whether to be adopted one-tailed or two-tailed test.

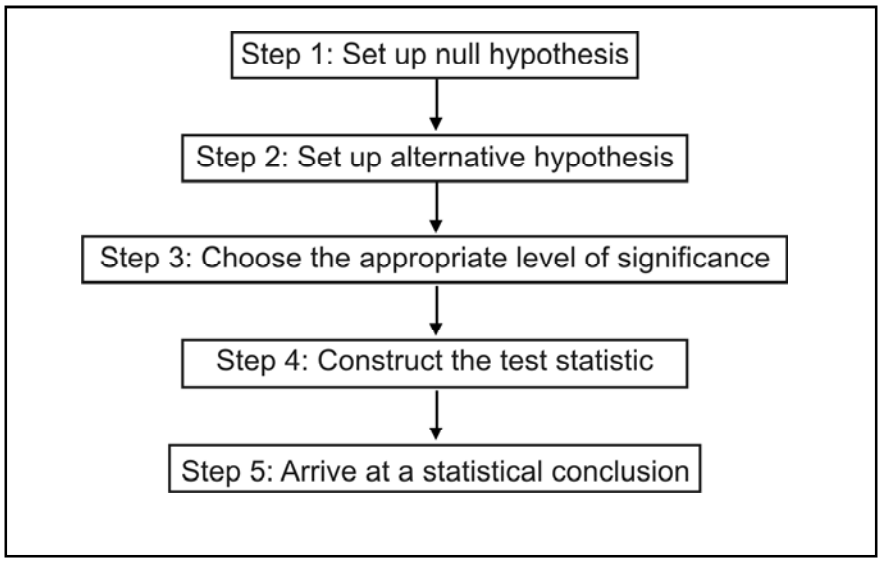
(3) To choose an appropriate level of significance. Commonly used levels of significance are: 0.01, 0.05 and 0.10.

(4) After setting up the null and alternative hypotheses, the researcher needs to select an appropriate statistical test that will be used for analyzing statistical data. Type, number and the level of data may provide a platform in the selection of the statistical test. Apart from these, the statistics used in the study (mean, proportion and variance etc.) must also be considered when a researcher selects an appropriate test, which can be used for testing of hypothesis to obtain the best results. Some of the widely used testing procedures are: z, t, F and χ^2.

(5) The final step in any hypothesis testing is to draw a statistical decision, which involves either to accept or to reject a null hypothesis. This depends on whether the calculated value of the test statistic falls in the acceptance region or in the rejection region at a given level of significance. In other words, if the computed value of Z (in modulus value) is less than the critical value Z_α then we say that it is not significant and null hypothesis may be accepted. On the other hand, if the computed value of Z (in modulus value) is greater than the critical value Z_α then we say that it is significant and null hypothesis is rejected at the level of significance α.

Therefore, various steps to be followed in testing of a statistical hypothesis can be summarized as shown in Figure 13.5

Figure 13.5: Steps in testing of a Statistical Hypothesis



Power of Test: Another way to evaluate the goodness of a statistical test is the complement of type II error, known as power of a test, which is defined as:

Power of a test = P (rejecting H_0 when it is false)

= P (rejecting H_0 when H_1 is true)

= 1- P (accepting H_0 when H_1 is true)

=1- P (type II error)

=1- β

Where β is the probability of type II error. Among a class of tests, the best is the one which has the maximum power for the same size.



CHECK YOUR PROGRESS

Q 1: What is meant by Hypothesis testing?.....

.....

Q 2: What is neant by Type I error?

.....

.....

Q 3: What is meant by Type II error?

.....

.....

13.7 TESTING OF HYPOTHESIS IN CASE OF LARGE SAMPLES

Here we shall develop the tests of significance in case of large samples. In practice, a sample is to be considered as large if the sample size is greater than or equal to 30 i.e., $n \geq 30$. Now since the assumptions made in case of large samples are not valid for small samples therefore the tests of significance used to deal with problems relating to large sample are different from the ones used for small samples.

The following are the assumptions made in testing of hypothesis relating to large samples:

(i) The population from which the sample is drawn is normally distributed. Consequently the sampling distribution of the sample statistic is also

normal.

(ii) Since sample size is large, due to central limit theorem, the value given by any random sample can be used in place of population value for calculating the standard error of the estimate.

13.7.1 Hypothesis Testing for Single Population Mean

Suppose a random sample of size $n(n \geq 30)$ with mean $\bar{x}$ is drawn from a population. We hold the claim that the population mean μ has the specified value μ_0.

Here we set up the null hypothesis $H_0 : \mu = \mu_0$. Against this hypothesis we may have any one of the following three alternative hypotheses

(i) $H_1 : \mu \neq \mu_0$ (Two-tailed test)

(ii) $H_1 : \mu > \mu_0$ (Right-tailed test)

iii) $H_1 : \mu < \mu_0$ (Left-tailed test)

Under the null hypothesis the test statistic to be constructed is:

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

where $\mu = \mu_0$, σ is the population standard deviation, n is the sample size and $\bar{x}$ is the sample mean.

In many real life situations, the population standard deviation remains unknown. To overcome such type of situations, in case of large sample ($n \geq 30$), the unknown population standard deviation σ can be replaced by a sample standard deviation (s).

For the comparison of the calculated value of Z with the critical value at the given level of significance, Table 13.1 is used and on the basis of the comparison decision to accept or reject the null hypothesis is taken as per the procedure discussed in 13.6.

We will now discuss the following examples.

Example 1: A business firm conducted a sample survey 15 years ago and claimed that the average household income of a particular geographic region is Rs. 10,000. Mr. Baruah, recently appointed as

vice president in the firm has expressed doubts about the accuracy of the data. In order to verify the claim, a random sample of 200 households is tested. It is found that the mean household income is Rs. 11,000. Assume that the population standard deviation of the household income is Rs. 1200. Can we support the claim at 0.05 level of significance?

Solution: Here first we set up the null hypothesis H_0 and alternative hypothesis H_1 as follows:

$$H_0 : \mu = 10{,}000$$

$$H_1 : \mu \neq 10{,}000$$

Given, $n = 200$, $\bar{x} = 11{,}000$, $\sigma = 1200$ and $\alpha = 0.05$

Since the sample size is large (≥ 30), therefore the Z test can be applied to test hypothesis.

Under the null hypothesis, the test statistic is given by

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$= \frac{11{,}000 - 10{,}000}{\frac{1200}{\sqrt{200}}} = 11.79$$

Therefore the calculated value of Z is 11.79

Again the critical value of Z at 5% level of significance for two-tailed test =1.96 Since the calculated value is Z greater than the critical value of Z, therefore the statistical decision is to reject the null hypothesis and conclude that the claim is not valid. Mr. Baruah's doubt about average household income was right.

Example 2: A company conducted a survey in the past to its consumers which reveals that the average household income is Rs. 20,000 in a particular region. Due to the rapid growth of development the company thinks that the average income may have changed. To verify this, a random sample of 50 households is selected and found that sample mean is Rs. 30,000 with a standard deviation of Rs. 15,000. Use 5% level of significance to verify the finding.

Solution: In the light of the problem, the null hypothesis H_0 and

alternative hypothesis H_1 are given as

$$H_0 : \mu = 20{,}000$$

$$H_1 : \mu \neq 20{,}000$$

Given, $n = 50$, $\bar{x} = 30{,}000$, $s = 15{,}000$ and $\alpha = 0.05$

Since the sample size is large (≥ 30), therefore the Z test can be applied to test hypothesis.

Under the null hypothesis, the test statistic is given by

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

(since the population S.D. σ is unknown, therefore it can be replaced by the sample S.D. s)

$$= \frac{30{,}000 - 20{,}000}{\frac{15{,}000}{\sqrt{50}}} = \frac{70.71}{15} = 4.714$$

Therefore the calculated value of Z is 4.714

Again the critical value of Z at 5% level of significance for two-tailed test =1.96 Since the calculated value Z is greater than the critical value of Z therefore the statistical decision is to reject the null hypothesis and conclude that average household income is significantly different from Rs. 20,000.

Example 3: The mean life of a sample of 80 fluorescent electric bulbs produced by a company is found to be 2000 hours with a standard deviation of 400 hours. Test the hypothesis that the mean life time of the bulbs produced in general is higher than the mean

life of 1900 hours at $\alpha = 0.01$ level of significance.

Solution: For the problem, let us take the null hypothesis that mean life of bulbs is not more than 1900 hours. Then the null hypothesis H_0 and alternative hypothesis H_1 will be as follows

$$H_0 : \mu \leq 1900$$

$$H_1 : \mu > 1900 \text{ (Right-tailed test)}$$

Given, $n = 80$, $\bar{x} = 2000$, $s = 400$ and $\alpha = 0.01$

Since the sample size is large (≥ 30), therefore the test can be applied to test hypothesis.

Under the null hypothesis, the test statistic is given by

$$Z = \frac{x - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

(since the population S.D. σ is unknown, therefore it can be replaced by the sample S.D. s)

$$= \frac{2000 - 1900}{\frac{400}{\sqrt{80}}}$$

$$= \frac{100}{44,72}$$

$$= 2.23$$

Therefore the calculated value of Z is 2.23

Again the critical value of Z at 1% level of significance for right-tailed test =2.33

Since the calculated value Z is less than the critical value of Z, therefore the statistical decision is to accept the null hypothesis. Hence we conclude that mean life of bulbs produced by the company is not more than 1900 hours.

13.7.2 Hypothesis Testing for Single Population Proportion

There are many areas where information that is highly dynamic in nature is generally expressed in terms of proportions or percentage such as, we often find that the market share of a company is 40% or 25% of the customers have switched from one brand to another brand or 10% of the items are defective. Quality defects, consumer preferences, market share etc. are some of the common areas where researchers may be interested to check the hypothesis whether such proportions have changed. For this a random sample of size n is selected from a large population possessing a particular attribute of interest (also termed as success) X then

$$p = \text{observed sample proportion of successes} = \frac{\mathrm{X}}{\mathrm{n}}$$

Here we are interested to test the hypothesis that the proportion P in the population has a specified value P_0.

We set up the null hypothesis $H_0 : P = P_0$

Depending upon the nature, the alternative hypothesis may be any one of the following:

(i) $P \neq P_0$ (Two-tailed test)
(ii) $P > P_0$ (Right-tailed test)
(iii) $P < P_0$ (Left-tailed test)

Since n is large, therefore under the null hypothesis, the test statistic to be constructed is

$$\mathrm{Z} = \frac{\mathrm{p} - \mathrm{P}}{\sqrt{\frac{\mathrm{PQ}}{\mathrm{n}}}} \sim \mathrm{N}(0,1)$$

where $P = P_0, Q = 1 - P = 1 - P_0$

To compare the calculated value of Z with the critical value at the given level of significance for the statistical decision, we use the Table 13.1.

Example 4: The production manager of a company that manufactures electric heaters claims that at least 10% of the heaters are defective. In order to test his believe, he takes a random sample of 100 heaters and finds that 12 heaters are defective. Use 5% level of significance to test the hypothesis.

Solution: Given $n = 100$, Number of defective heaters $= 12$

$p =$ sample proportion of defective heaters $= \frac{12}{100} = 0.12$

Null hypothesis $H_0 : P = 0.10$

Alternative hypothesis $H_1 : P \neq 0.10$

Since the sample size is large (≥ 30), therefore the Z test can be applied to test hypothesis.

Under the null hypothesis, the test statistic is given by

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$

$$= \frac{0.12 - 0.10}{\sqrt{\frac{0.10 \times 0.90}{100}}} \quad (\because Q = 1 - P = 1 - 0.10 = 0.90)$$

$$= \frac{0.02}{0.03} = 0.67$$

$\therefore$ The calculated value of $Z = 0.67$

The critical value of Z for two-tailed test at 5% level of significance=1.96 Since the calculated value is less than the critical value (0.67<1.96) therefore the statistical decision leads to acceptance of null hypothesis and conclude that the production manager's claim that at least 10% of the products are defective is valid. Consequently, the manufacturer can plan a strategy considering the fact that 10% of the products may be defective.

Example 5: A manufacturer of LCD TV claims that it is becoming quite popular in Guwahati, and that about 5% homes are having LCD TV. However, a dealer of conventional TVs claims that the percentage of homes with LCD TV is less than 5%. A sample of 500 household is surveyed, and it is found that only 18 households have LCD TV. Test at 1% level of significance whether the claim of the company is tenable.

Solution: Given, Given $n = 400$, Number of households having LCD TV

$p =$ sample proportion of LCD TV owners $= \frac{18}{500} = 0.045$

Null hypothesis $H_0 : P = 0.05$

Alternative hypothesis $H_1 : P < 0.05$

Since the sample size is large (≥ 30), therefore the Z test can be applied to test hypothesis.

Under the null hypothesis, the test statistic is given by:

$$Z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$

$$= \frac{0.045 - 0.05}{\sqrt{\frac{0.05 \times 0.95}{500}}} \quad (\because Q = 1 - P = 1 - 0.05 = 0.95)$$

$$= \frac{-0.005}{0.0097}$$

$$= -0.515$$

$$\therefore \quad |Z| = 0.515$$

Therefore the calculated value of $|Z| = 0.515$

The critical value of Z for one-tailed test at 1% level of significance = 2.33 Since the calculated value is less than the critical value (0.515 < 2.33) therefore the statistical decision leads to the acceptance of null hypothesis and conclude that the manufacturer's claim that about 5% homes in Guwahati are having LCD TV is valid.

Example 6: The manager of a workshop claimed that at least 95%

of the equipments which he supplied to a factory conformed to specifications. To verify his claim a sample of 200 pieces of equipments was selected and revealed that 18 were faulty. Test his claim at a significance level of 0.05.

Solution: Given n=200,

$P =$ Proportion of pieces conforming to specification in the lot = 95% =0.95

$Q =$ Proportion of pieces not conforming to specification in the lot = $1 - P = 0.05$

Sample proportion $\mathrm{p} = \frac{200 - 18}{200} = 0.91$

We take the null hypothesis that the proportion of pieces conforming to specification is 95% i.e., $H_0 : P = 0.95$

Alternative hypothesis $H_1 : P < 0.10$ (left-tailed test)

Since the sample size is large (≥ 30), therefore the Z test can be applied to test hypothesis.

Under the null hypothesis, the test statistic is given by

$$Z = \frac{\mathrm{p} - \mathrm{P}}{\sqrt{\frac{\mathrm{PQ}}{\mathrm{n}}}}$$

$$= \frac{0.91 - 0.95}{\sqrt{\frac{0.95 \times 0.91}{200}}}$$

$$= \frac{-0.04}{0.015}$$

$$= -2.6$$

$$\therefore |Z| = 2.6$$

The critical value of $|Z|$ at $\alpha = 0.05$ for one-tailed test = 1.645.

Since the calculated value of $|Z|$ is greater than the critical value

(i.e., 2.6 >1.645) therefore we reject the null hypothesis at 5% level and conclude that the claim of the manager is not valid.

[Alternatively, without taking the modulus we may compare the calculated value of Z with the critical value of $Z = -1.645$ at $\alpha = 0.05$ for left-tailed test. Since $-2.6 < -1.645$, the null hypothesis is to be rejected and thus both types of comparison lead us to the same decision.]



CHECK YOUR PROGRESS

Q 4: State whether the following statements are true or false:

(i) A tentative assumption about a population parameter is called the null hypothesis.

(ii) The equality part of the expression (either $\geq, \leq,$ or $=$) always appears in the null hypothesis.

(iii) The estimated variance of sampling distribution of a statistic is called standard error.

(iv) The test statistic to test $\mu = \mu_0$ for large sample is z -test.

(v) The usual notation for the standard error of the sampling distribution is σ/n.

(vi) Suppose for a particular hypothesis $\alpha = 0.05$ and $\beta = 0.10$. Then the power of this hypothesis test is 0.90.

(vii) If 2.81 is the critical value of test-statistic z, then the significance level of the test is 0.05.

(viii) The large sample test for testing $P = P_0$ for normal population is t-test.

(viii) For test of hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$, the critical region at $\alpha = 0.05$ and $n \geq 30$, is $z > 1.96$.

(ix) When using the sample proportion $\bar{p}$ to test the hypothesis, the standard error of $\bar{p}$ is $\sqrt{\bar{p}\bar{q}/n}$.

(x) Sampling distribution will be approximately normal if sample size is small.

P – Value Approach for Hypothesis Testing: This approach is applied in hypothesis testing for large samples and sometimes referred to as the observed level of significance. This approach is

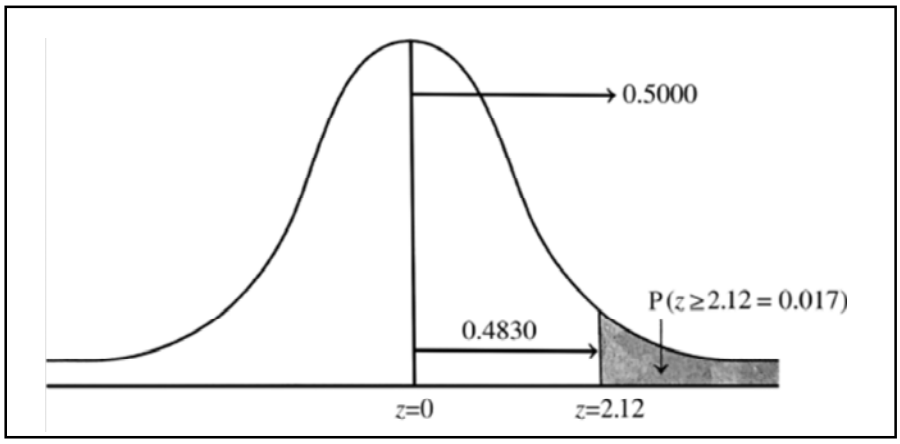
very advantageous especially after the introduction of many statistical software packages. The $P_$ value defines the smallest value of α for which the null hypothesis can be rejected. The decision rule for accepting or rejecting a null hypothesis based on the $P_$ value as discussed below:

Reject the null hypothesis H_0 when the value is $< \alpha$, otherwise, accept the null hypothesis H_0

For example, suppose a researcher conducting a hypothesis test with rejection region on the right tail of the distribution obtains an observed test statistic value of 2.12. From the normal table, the corresponding probability area for z value 2.12 is 0.4830. Therefore, the probability of getting a value greater than or equal to 2.12 is 5.0000-0.4830=0.017 (as shown in the following figure)

In this case, the p value is 0.017. For α =0.05 and α =0.01, this value falls under the rejection region (α =0.05, 0.017< 0.05 and α =0.01, 0.017< 0.01), so the null hypothesis will be rejected at 0.05 0.01 levels of significance. At α =0.01, the researcher cannot reject the null hypothesis for the value of equal to 0.017 because α =0.01< 0.017.

Figure 13.6: Probability of p (z is greater than or equal to 2.12)



The procedure of calculating the p value for a two-tailed test is slightly different. In a two-tailed test, the rejection region falls in both the tails of the normal distribution. For example, at α =0.05, the rejection region is located in both the tail areas of the distribution

in terms of 0.025% (0.025 +0.025=0.05) area in both the tails of the distribution. The researcher compares this value to the computed value for accepting or rejecting the null hypothesis. For this comparison, instead of splitting the value, we double the value and then compare that value with the value. In the above mentioned example, the researcher conducted a hypothesis test with rejection region on both the tails of the distribution and obtained the observed test statistic as 2.12 and corresponding value as 0.017. So, instead of splitting value in two parts, the researcher should double the value, i.e., (0.017×2=0.034). So, for a two-tailed test, this value (0.034) is compared with the values 0.05 and 0.01. Hence, at α =0.05 and α =0.01 the null hypothesis will be rejected but at α =0.01 the null hypothesis will be accepted (0.01< 0.034).

Example 7: A marketing research firm conducted a survey 10 years ago and observed that the average household income of a particular geographic region is Rs. 10,000. Mr. Gupta, who has recently joined the firm as a vice president has expressed doubts about the accuracy of the data. For verifying the data, the firm has decided to take a random sample of 200 households that yield a sample mean (for household income) of Rs. 10,200. Assume that the population standard of the household income is Rs. 1200. Use the p -value approach to verify Mr. Gupta's doubts. (Use $\alpha = 0.01$ as the level of significance.)

Solution: We set up the null and alternative hypothesis given by

Null hypothesis $H_0 : \mu = 10{,}000$

Alternative hypothesis $H_1 : \mu \neq 10{,}000$

Since the sample size is large (≥ 30), therefore the Z test can be applied to test hypothesis.

Under the null hypothesis, the test statistic is given by

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$= \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

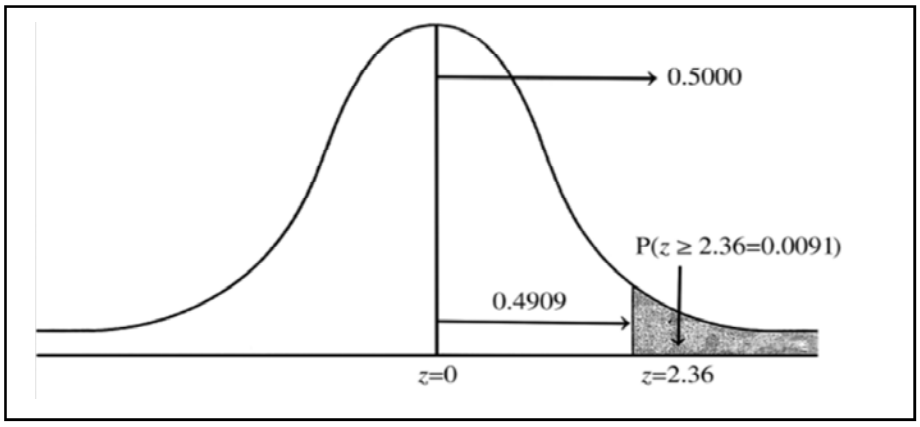
(since the population S.D. σ is unknown, therefore it can be replaced by the sample S.D. s)

$$= \frac{10{,}200 - 10{,}000}{\frac{1200}{\sqrt{200}}} = \frac{200}{84.85} = 2.36$$

Therefore the observed test statistic is computed as 2.36.

From the normal table, the corresponding probability area for Z value = 2.36 is 0.4909. So, the probability of obtaining a Z value greater than or equal to 2.36 is 0.5000-0.4909=0.0091 (as obtained from the following figure).

Figure 13.7: Probability of p (z is greater than or equal to 2.36)



For a two-tailed test, the obtained area value 0.0091 is to be multiplied by 2 and thus producing the value 0.0091×2=0.0182. Now since, $0.01 < 0.0182$, therefore we may accept the null hypothesis.

13.7.3 Hypothesis Testing for Difference between Two Population Means

In the real-world business environment, business researchers often encounter situations where they need to compare two means taken from two different samples drawn from two populations. On the

basis of the samples researchers make an inference about the difference in the means of the two populations. For example, suppose a researcher is interested to conduct a survey to study the consumer satisfaction for a particular brand of product in two different cities, Guwahati and Dibrugarh. In order to accomplish this, the researcher selects two different random samples from the two cities taken in the study. By using the procedures of testing of hypothesis, finally, the researcher draws a statistical conclusion regarding the consumer satisfaction in the two cities. Now, let $\overline{x_1}$ and $\overline{x_2}$ be the means of two large random samples of sizes n_1 and n_2 drawn from two populations with means μ_1 and μ_2 and standard deviations σ_1 and σ_2. We want to test the hypothesis whether the means of the two populations are equal or if there exists any significant difference between the population means.

Thus the null hypothesis is $H_0 : \mu_1 = \mu_2$

The alternative hypothesis is $H_1 : \mu_1 \neq \mu_2$

Assuming the null hypothesis to be true, the test statistic

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

The decision rule is same as discussed in the earlier units.

Example 8: An investigation of the relative merits of two kinds of flashlight batteries showed that a random sample of 100 batteries of brand A lasted on the average for 35.5 hours with a standard deviation of 1.8 hours, while a random sample of 80 batteries of brand B lasted on the average 36.8 hours with a standard deviation of 1.5 hours. Do the two brands differ significantly in quality? Use $\alpha = 0.05.$

Solution: Let μ_1 and μ_2 denote the means of two populations.

Let us set up the null hypothesis that the two brands do not differ significantly in quality, i.e., $H_0 : \mu_1 = \mu_2$

and an alternative hypothesis $H_1 : \mu_1 \neq \mu_2$

Since the sample size is large (≥ 30), therefore the Z test can be applied to test hypothesis.

Under the null hypothesis, the test statistic is given by

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

$$= \frac{36.5 - 36.8}{\sqrt{\frac{3.24}{100} + \frac{2.25}{80}}} = \frac{-0.3}{\sqrt{0.0324 + 0.0281}} = -1.2195$$

$$\therefore |Z| = 1.2195$$

The critical value of $|Z|$ for two-tailed test at $\alpha = 0.05$ is 1.96.

Since the calculated value of $|Z|$ is less than the critical value of $|Z|$ therefore we may accept the null hypothesis and conclude that there is no significant difference between the average life times of the two brands of batteries.

Example 9: The means of two large samples of size 1000 and 2000 are found to be 67.5 and 68.0 respectively. Test the equality of the two populations each with S.D. 2.5.

Solution: Let us set up null and alternative hypothesis as:

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

Where μ_1 is the mean of the first population and μ_2 is the mean of the second population.

Since the sample size is large (≥ 30), therefore the Z test can be applied to test hypothesis.

Under the null hypothesis, the test statistic is given by

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

$$= \frac{67.5 - 68.0}{\sqrt{\frac{6.25}{1000} + \frac{6.25}{2000}}} = \frac{-0.5}{\sqrt{0.00625 + 0.00313}} = -5.16$$

$$\therefore |Z| = 5.16$$

Since the calculated value of $|Z|$ is greater than the critical value of $|Z|$ therefore we reject the null hypothesis and conclude that there is significant difference between population means.



CHECK YOUR PROGRESS

Q 5: A sample of 40 observations is selected from one population with S.D. (σ) 5. The sample mean is 102. A sample of 50 observations is selected from a second population with S.D. (σ) 6. The sample mean is 99. Conduct the following test of hypothesis (use $\alpha = 0.05$)

(a) Is this a one tailed or a two-tailed test? (b) State the decision rule. (c) Compute the value of the test statistic. (d) What is your decision regarding H_0?

13.7.4 Hypothesis Testing for Difference between Two Population Proportions

At times, we come across certain situations where we are interested to study whether or not there exist significant differences between the proportions of two groups of, say, consumer in respect of a certain activity. Let p_1 and p_2 be the proportions of a certain attribute in two large random samples of sizes n_1 and n_2 drawn from two populations having proportions P_1 and P_2. On the basis of the samples, we want to test whether the two population proportions are equal or if there exist any significant difference between the

population proportions. Thus we will have null and alternative hypotheses as:

$$H_0 : P_1 = P_2$$
$$H_1 : P_1 \neq P_2$$

Assuming the null hypothesis to be true, the test statistic

$$Z = \frac{P_1 - P_2}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}} \sim N(0,1)$$

where $Q_1 = 1 - P_1, \quad Q_2 = 1 - P_2.$

The procedure of taking decision is same as discussed earlier.

Note: (i) If P_1 and P_2 remains unknown, then they can be estimated by the corresponding sample proportions p_1 and p_2.

(ii) Sometimes, population proportion can be estimated by pooling the two sample proportions p_1 and p_2 into a single proportion p as given

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

In this case, the test statistic Z can be restated as

$$Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad q = 1 - p$$

Example 10: In two populated states there are 30% and 25% respectively of fair haired people. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations?

Solution: Given $n_1 = 1200, \quad n_2 = 900$

$$P_1 = 30\% = 0.30 \Rightarrow Q_1 = 0.70; \quad P_2 = 25\% = 0.25 \Rightarrow Q_2 = 0.75$$

Now we set up the null hypothesis H_0 that the difference in population proportions is likely to be hidden in samples i.e.,

$$H_0 : P_1 = P_2$$

And the alternative hypothesis is

$$H_1 : P_1 \neq P_2$$

Since the samples are large, under H_0 the test statistic Z is given by

$$Z = \frac{P_1 - P_2}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}} \sim N(0,1)$$

$$= \frac{0.30 - 0.25}{\sqrt{\frac{0.30 \times 0.70}{1200} + \frac{0.25 \times 0.75}{900}}}$$

$$= \frac{0.05}{\sqrt{0.000175 + 0.000208}} = \frac{0.05}{0.0196} = 2.55$$

Therefore the calculated value of Z =2.55

The critical value of Z at 5% level for two tailed test=1.96

Since the calculated value of Z > the critical value of Z, We reject the null hypothesis and hence, conclude that the difference in population proportions is not likely to be hidden in these samples.

Example 11: A Company has the head office at Calcutta and a branch at Guwahati. The personnel director wanted to assess if the workers at the two places would like the introduction of a new plan of work to enhance the quality of services and a survey was conducted for the purpose. Out of a sample of 500 workers at Calcutta 62% favoured the new plan. At Guwahati out of a sample of 400 workers 41% were against the new plan. Is there a significant difference between the two groups in their in their attitude towards the new $n_2 = 400, \quad p_2 = 59\% = 0.59 \Rightarrow q_2 = 1 - 0.59 = 0.41$ plan at 5% level?

Solution: Given

$$n_1 = 500, \quad p_1 = 62\% = 0.62 \Rightarrow q_1 = 1 - 0.62 = 0.38$$

Now we set up the null hypothesis that the population proportion of workers favouring the new plan in Calcutta and Guwahati is same, i.e. $H_0 : P_1 = P_2$,

The alternative hypothesis is

Since the samples are large, under the test statistic is given by

$$Z = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

Where $p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}, \quad q = 1 - p$

$$Z = \frac{0.62 - 0.59}{\sqrt{0.007 \times 0.393\left(\frac{1}{500} + \frac{1}{400}\right)}},$$

$$= \frac{0.03}{0.0327} = 0.917$$

Therefore the calculated value of Z =0.917

The critical value of Z at 5% level for two tailed test=1.96

Since the calculated value of $Z <$ the critical value of , We may accept the null hypothesis and hence, conclude that there is no significant difference between the two groups in their attitude towards the new plan.



13.9 LET US SUM UP

- Hypothesis testing is a well-defined procedure that helps us to decide objectively whether to accept or reject the hypothesis based on the information available from the sample. Testing of hypothesis involves five different steps.
- In case of testing of statistical hypothesis, there are two types of hypothesis tests namely two-tailed test and one-tailed test. Two-tailed tests contain the rejection region on both the tails of the sampling distribution of a test statistic. As different from a two-tailed test, a one-tailed test contains the rejection region on one tail of the sampling

distribution of a test statistic.

- The testing of hypothesis involves two types of errors: Type I error and Type II error. A Type I error is committed by rejecting a null hypothesis when it is true. The possibility of committing Type I error is termed as level of significance and is denoted by α. On the other hand, a type II error is committed by accepting a null hypothesis when it is false. The likelihood of committing Type II error is denoted by β.
- Hypothesis testing can be performed by applying the approaches: the $z-$value approach and $p-$value approach. For testing hypothesis about a single population mean, z formula can be used if the sample size is large $(n \geq 30)$ for any population with known population standard deviation σ. But in practical situations if σ is not available then it can be estimated by a sample standard deviation s. The p value defines the smallest value of for which the null hypothesis can be rejected.
- Hypothesis can also be performed for single population proportion as well as for the difference in two population proportions. In case of two random samples, on the basis of the difference in sample proportions, a researcher can estimate the difference in population proportions.



13.10 FURTHER READING

- Gupta S.C. (1994). Fundamentals of statistics. New Delhi, Himalayan Publishing House.
- Sharma, J.K. (2007). Business Statistics. New Delhi. Pearson Education Ltd.
- Hazarika, P.L. (2016). Essential Statistics For Economics And Business Studies. New Delhi. Akansha Publishing House.
- Lind, D.A., Marshal, W.G., Wathen, S.A. (2009) Statistical Techniques in Business and Economics. New Delhi. Tata McGraw Hill Education Private Limited.

13.11 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: Hypothesis testing is the process of testing a statement or belief about a population parameter by the use of information collected from a sample (s).

Ans to Q No 2: Type I error is the probability of rejecting the null hypothesis when it is true. It is denoted by Greek letter α .

Ans to Q No 3: Type II error is the probability of accepting the null hypothesis when it is false. It is denoted by Greek letter β .

Ans to Q No 4: (i) T, (ii) T, (iii) F, (iv) T, (v) F, (vi) T, (vii) F, (viii) F, (ix) T, (x) F.

Ans to Q No 5: (a) Two-tailed test, (b) Reject H_0 $|z| > 1.96$,

(c) $$z = \frac{102 - 99}{\sqrt{\frac{5^2}{40} + \frac{6^2}{50}}} = 2.59$$

(d) Reject H_0

13.12 MODEL QUESTIONS

Q 1: Discuss the various steps involved in testing of hypothesis.

Q 2: What do you understand by null hypothesis, alternative hypothesis and level of significance?

Q 3: Define Type I error and Type II error.

Q 4: What is meant by one-tailed test and two-tailed test?

Q 5: The mean breaking strength of the cables supplied by a manufacturer is 1800 with a standard deviation of 100. By a new technique in the manufacturing process it is claimed that the breaking strength of the cables has increased. In order to test this claim a sample of 50 cables is tested. It is found that the mean breaking strength is 1850. Can we support the claim at 1% level of significance?

Q 6: The MacBurger restaurant chain claims that the mean waiting time of customers is 3 minutes with a population standard deviation of 1 minute. The quality- assurance department found in a sample of 50

customers at the Warren Road MacBurger that the mean waiting time was 2.75 minutes. At the 5% significance level, can we conclude that the mean waiting time is less than 3 minutes?

Q 7: A company manufacturing a certain type breakfast cereal claims that 60% of all housewives prefer that type to any other. A random sample of 300 housewives contains 165 who do prefer that type. At 5% level of significance, test the claim of the company.

Q 8: A man buys 50 electric bulbs of 'Philips' and 50 electric bulbs of 'HMT'. He finds that 'Philips' gave an average life of 1500 hours with a standard deviation of 60 hours and 'HMT' bulbs gave an average life of 1512 hours with a standard deviation of 80 hours. Is there a significant difference in the mean life of the two makes of bulbes?

Q 9: 500 units from a factory are inspected and 12 are found to be defective, 800 units from another factory are inspected and 12 are found to be defective. Can it be concluded that at 5% level of significance production at the second factory is better than in first factory?

*** ***** ***

UNIT 14: TEST OF SIGNIFICANCE

UNIT STRUCTURE

14.1 Learning Objectives

14.2 Introduction

14.3 The Chi-Square (χ^2) Distribution

14.4 Applications of χ^2 Test

- 14.4.1 χ^2 Test of Goodness of Fit
- 14.4.2 χ^2 Test for Independence of Attributes
- 14.4.3 χ^2 Test for a Specified Population Variance
- 14.4.4 χ^2 Test of Homogeneity

14.5 Testing of Hypothesis in case of Small Samples (t-Test)

- 14.5.1 Characteristics of t-distribution
- 14.5.2 Applications of t-distribution
- 14.5.3 Hypothesis Testing for Single Population Mean
- 14.5.4 Hypothesis Testing for Difference between Two Population Means
- 14.5.5 Paired t-test for Difference between Two Population Means

14.6 Let Us Sum Up

14.7 Further Reading

14.8 Answer to Check Your Progress

14.9 Model Questions

14.1 LEARNING OBJECTIVES

After going through this unit you will be able to

- introduce the Chi-square distribution for the comparison of more than two means or proportions
- use the Chi-square distribution to test whether two classifications of the same data are independent of each other
- use the Chi-square distribution for testing the hypothesis about a single population variance
- use t-test in small samples.

14.2 INTRODUCTION

In the previous unit we have learnt some statistical tests (using data

from either one or two samples) that are used to test the hypothesis about the population parameters, with a number of business related examples. These tests are based on the assumption that the samples are drawn from a normally distributed population or, more appropriately, that the sample means are normally distributed. Such tests are known as parametric tests as the testing procedure involves the assumption about the type of population of parameters. On the other hand, in some cases, the researcher may not be sure of whether the population distribution is normal. The statistical tests that do not require prior knowledge about the population are termed as non-parametric tests. Chi-square test is one of the important non-parametric methods.

A non-parametric method, also known as distribution free test satisfies at least one of the following criteria:

(i) The method does not take into consideration any population parameter.

(ii) It is applied only on categorical data that are non-numerical and frequency counts of categories for one or more variables.

(iii) It does not depend on the form of the underlying population distribution, in particular, the requirement of normality.

In this unit we will focus on χ^2 (Chi-square) test. This test enables us to test whether more than two population proportions are equal. In fact, it allows to do a lot more than just test for the equality of several proportions. If we classify a population into several categories with respect to two attributes (such as age and job performance), we can then employ a chi-square test to test whether the two attributes are independent of each other.

14.3 THE CHI-SQUARE (χ^2) DISTRIBUTION

The symbol χ stands for the Greek letter 'chi'. This test was first developed by Karl Pearson in testing of hypothesis in 1900. The sampling distribution of χ^2 is called χ^2 distribution. This distribution has only one parameter called the degrees of freedom (df) and the shape of the curve depends on the degrees of freedom. The distribution is skewed to the right. Being a square quantity, χ^2 distributi3on can never be negative. In fact, χ^2 is a continuous probability distribution with range 0 to ∞ (Figure 14.1)

The probability density function of distribution is given by

$$f(\chi^2) = C(\chi^2)^{\frac{\nu}{2}-1} e^{-\frac{\chi^2}{2}}$$

where ν is is the degrees of freedom, C is a constant depending upon the degrees of freedom, and $e = 2.71828.$

χ^2 test statistic can be defined as below:

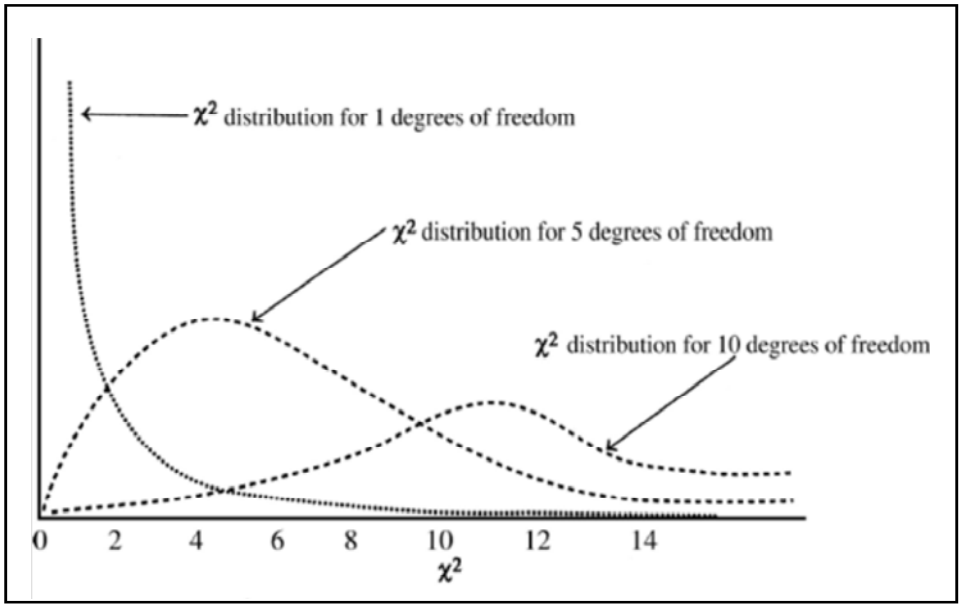
$$\chi^2 = \sum \frac{(O-E)^2}{E} \text{ with } df = k - 1 - c$$

where O is the observed frequency, E the expected frequency, k the number of categories, and c the number of parameters being estimated from the sample data.

The computed value of χ^2 is compared with the critical value of χ^2 at a specified level of significance and degrees of freedom. The decision to accept or reject the null hypothesis is given by

If $\chi^2_{cal} < \chi^2_{critical}$ then we may accept the null hypothesis, otherwise we reject the null hypothesis.

Figure 14.1: χ^2 Distribution with 1, 5 and 10 Degrees of Freedom



Properties of Chi-square test:

The following properties are useful while using χ^2 test statistic to analyze its sampling distribution:

(i) The density function of χ^2 does not involve any parameter and the shape of the curve depends upon the degrees of freedom only. The distribution becomes more and more symmetrical as the degrees of freedom increase and tends to normal distribution for high value of degrees of freedom ($\nu \geq 30$).

(ii) Mean = Degrees of freedom = ν, mode= $\chi^2 - 2$ and variance= 2ν

(iii) The values of χ^2 lie between 0 and ∞.

(iv) If $x_1, x_2, ..., x_n$ be a random sample drawn from a normal population with mean μ and variance σ^2, then

(a) $\sum_{i=1}^{n} (x_i - \mu)^2 / \sigma^2$ follows χ^2 -distribution with n degrees of freedom.

(b) $\sum_{i=1}^{n} (x_i - \bar{x})^2 / \sigma^2$ follows χ^2 -distribution with degrees of freedom $n-1$, $\bar{x}$ being the sample mean.

(v) For large degrees of freedom (ν), the standard variate

$$Z = \frac{\chi^2 - E(\chi^2)}{S.E.(\chi^2)} = \frac{\chi^2 - \nu}{\sqrt{2\nu}}$$

may approximately be considered as a standard normal variate.

Conditions for applications of Chi-square test:

The following conditions need to be satisfied before applying χ^2 as a test statistic for hypothesis testing:

(i) In a contingency table, an expected frequency of less than 5 in a cell is less than the frequency required to apply the test. In such cases we apply the χ^2 technique of 'pooling' which consists in adding the

may approximately be considered as a standard normal variate.

Conditions for applications of Chi-square test:

The following conditions need to be satisfied before applying χ^2 as a test statistic for hypothesis testing:

(i) In a contingency table, an expected frequency of less than 5 in a cell is less than the frequency required to apply the test. In such cases we apply the technique of 'pooling' which consists in adding the frequencies which are less than 5 with the preceding or succeeding frequency, so that the resulting sum is 5 or more and adjust for the degrees of freedom accordingly.

(ii) The sample should consist of at least 50 observations and should be drawn randomly from the population.

(iii) The data should be expressed in original units, rather than in percentage or ratio form.

(iv) The constraints on the cell frequencies, if any, should be linear such as $\sum O = \sum E = N$, the total frequency.

14.4 APPLICATIONS OF χ^2 TEST

The Chi-square test is not based on assumption of normality and can be used to both large and small samples. Chi-square distribution is a widely used since it can be employed in situations in which the normal distribution cannot be applied. Chi-square distribution has a number of applications, some of which are enumerated as below:

(i) to test goodness of fit,

(ii) to test the independence of attributes,

(iii) to test if the population has a specified value of standard deviation or variance.

(iv) to test if two or more populations are homogenous with respect to some characteristics of interest.

Critical Values of χ^2 :

From the χ^2 table (given in the Appendix) it may be observed that that the tabulated (critical) values of χ^2 increase as the degrees of freedom increases and the level of significance decreases.

Let us define the symbol $\chi^2_n(\alpha)$ as the value of χ^2 for $n\ d.f.$ such that the area to the right of this point is α, i.e.,

$$P\left[\chi^2 > \chi^2_n(\alpha)\right] = \alpha$$

is the rejection region. This is shown in the Figure 14.2.

Figure 14.2: Critical Value and Rejection region of χ^2 Statistic

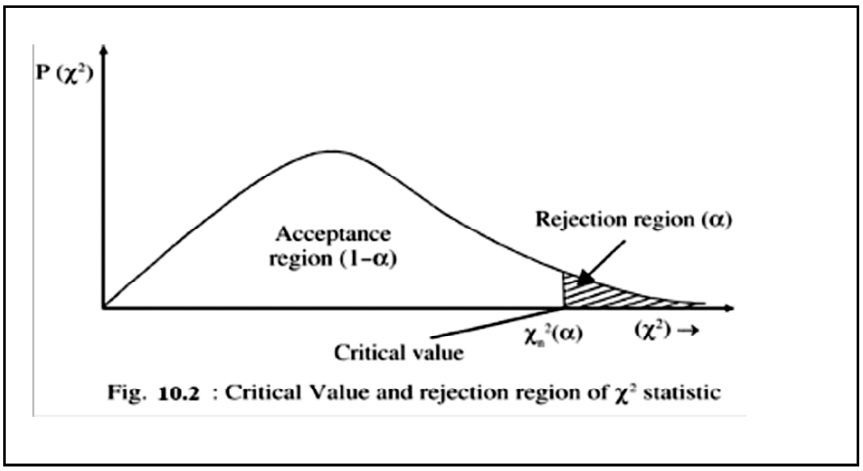


Fig. 10.2 : Critical Value and rejection region of χ^2 statistic



CHECK YOUR PROGRESS

Q 1. State whether the following statements are true or false:

(i) The critical value of a statistic is the value which cuts off the region for the rejection of null hypothesis.

(ii) The χ^2-test requires that the sampling distribution should be normally distributed.

(iii) If null hypothesis is accepted, then it is concluded that data analysis did not produce significant results.

(iv) In order to reject the null hypothesis $\chi^2_{cal} > \chi^2_{table}$

(v) The sampling distribution is a family of curves, the distribution of which varies with the degrees of freedom.

Now we discuss various applications of χ^2.

14.4.1 χ^2 Test of Goodness of Fit

χ^2 test enables us to ascertain whether the known probability distributions such as Binomial, Poisson and Normal distributions fit with an observed sample observation. Thus χ^2 test for goodness of fit, a statistical test, provides a platform that can be used to ascertain whether theoretical probability distributions coincide with empirical sample distributions. This test compares the theoretical (expected) frequencies with the observed (actual) frequencies to obtain the deviation between theoretical and observed frequencies. For the application of χ^2 test, first a theoretical distribution is hypothesized for a given population and then the test is carried out to make sure whether or not the sample data could come from the population of interest with the hypothesized theoretical distribution. The steps used for hypothesis testing can also be performed using the χ^2 goodness of fit test.

Steps to conduct χ^2 test for goodness of fit: The general steps to conduct a goodness of fit test for any hypothesized population distribution are summarized as below:

Step 1: Set up the null and alternative hypotheses as

Null hypothesis H_0 : No difference between the observed and expected sets of frequencies.

Alternative hypothesis H_1 : There is a difference.

Step 2: Calculate expected frequencies (E values) in each category by multiplying the category probability by the sample size ().

Step 3: Compute the test statistic

$$\chi^2 = \sum \frac{(\mathrm{O}-\mathrm{E})^2}{\mathrm{E}}$$

where O is the observed frequency, E the expected frequency.

Step 4: Using a specified level of significance α and $d.f. = n-1$ provided that the number of expected frequencies are 5 or more for all categories, determine the critical value of χ^2.

Step 5: Compare the computed value of χ^2 with the critical value

of χ^2 at a specified level of significance and degrees of freedom. The decision to accept or reject the null hypothesis is given by

If χ^2_{cal} is less than the critical value $\chi^2_{\alpha,n-1}$ then we may accept the null hypothesis, otherwise we reject the null hypothesis.

Example 1: A company is concerned about the increasing violent altercations between its employees. The number of violent incidents recorded by the management during six randomly selected months is given by

Months	Number of violent incidents
Jan	55
Feb	65
Mar	68
Apr	72
May	80
Jun	85

Test whether the data fits a uniform distribution. (Use $\alpha = 0.05$)

Solution: Let us set up null and alternative hypotheses as

Null hypothesis H_0 : Number of violent altercations are uniformly distributed over the months.

Alternative hypothesis H_1 : Number of violent altercations are not uniformly distributed over the months.

In order to test the null hypothesis, the appropriate test statistic is given by

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where O is the observed frequency, E the expected frequency.

Now the expected frequencies will be calculated by dividing total observed frequencies by number of

months. Hence, the expected frequency $\frac{\sum O}{6} = \frac{420}{6} = 70.$

Now to compute the test statistic we prepare the following table:

Months	O	E	$\frac{(O-E)^2}{E}$
Jan	55	70	3.2142
Feb	65	70	0.3571
Mar	68	70	0.0571
Apr	72	70	0.0571
May	80	70	0.9142
Jun	85	70	2.0571
Total	420		6.65

Thus $\chi^2 = 6.65$

$\therefore$ The calculated value of χ^2 is 6.65

The degrees of freedom= $n-1=6-1=5$

The critical value of χ^2 at 5% level of significance for 5 degrees of freedom i.e., $\chi^2_{0.05,5} = 11.07$

Since the calculated value of is less than the critical value , therefore we may accept the null hypothesis and conclude that the number of violent altercations is uniformly distributed over the months.

Example 2: A brand manager is concerned that her brand's share may be unevenly distributed throughout the country. In a survey in which the country was divided into four geographical regions, a random sampling of 100 consumers in each region was surveyed, with the following information:

	Region				
	NE	NW	SE	SW	Total
Purchase the brand	40	55	45	50	190
Do not purchase	60	45	55	50	210
Total	100	100	100	100	400

Develop a table of observed and expected frequencies for this problem.

(i) Calculate the sample Chi-square value.

(ii) State the null and alternative hypothesis.

(iii) At $\alpha = 0.05$, test whether brand share is the same across the four regions.

Solution: To calculate the expected frequencies for the corresponding observed frequencies, we apply the following formula

(Row total × Column total)/Grand total.

The calculations of expected frequencies for the corresponding observed frequencies are given below:

		O	E	$O-E$	$(O-E)^2$	$(O-E)^2/E$
Row 1	NE	40	47.5	-7.5	56.25	1.18
	NW	55	47.5	7.5	56.25	1.18
	SE	45	47.5	-2.5	6.25	0.13
	SW	50	47.5	2.5	6.25	0.13
Row 2	NE	60	52.5	7.5	56.25	1.07
	NW	45	52.5	-7.5	56.25	1.07
	SE	55	52.5	2.5	6.25	0.12
	SW	50	52.5	-2.5	6.25	0.12
Total						5.00

(i) The sample Chi-square statistic is given by

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where O is the observed frequency, E the expected frequency.

$\therefore$ From the above table we have,

$$\chi^2 = 5.00$$

(ii) We state the null and alternative hypotheses as:

Null hypothesis H_0 : The brand share is evenly distributed.

Alternative hypothesis H_1 : The brand share is not evenly distributed.

(iii) The calculated value of χ^2 is 5.00

The degrees of freedom = $(r-1)(c-1) = (2-1)(4-1) = 3$

The critical value of χ^2 at 5% level of significance for 3 degrees of freedom i.e., $\chi^2_{0.05,3} = 7.815$

Since the calculated value of is less than the critical value, therefore we may accept the null hypothesis and conclude that the brand share is evenly distributed in all the four regions of the country.

Example 3: A survey of 800 families with 4 children each revealed the following distribution:

Test whether the data are consistent with the hypothesis that the binomial law holds and the chance of a male birth is equal to that of a female birth.

Solution: Here we set up the null hypothesis that the data are consistent with the binomial law of equal probability for male and female births, i.e.,

$p = q = \frac{1}{2}$, (where p denotes the probability of a male birth.)

Using the binomial probability law, the expected frequency $f(x)$ of x male births is given by:

Number of births		Number of families
Male	Female	
0	4	32
1	3	178
2	2	290
3	1	236
4	0	64

$f(x) = N.p(x) = N \times {}^{n}C_{x}\, p^{x}q^{n-x}$, where

$N = 800,\ n = 4,\ \ x = 0, 1, 2, 3, 4.$

$$= 800 \times {}^{4}C_{x} \left(\frac{1}{2}\right)^{4}$$

$$= 50 \times {}^{4}C_{x}$$

……(*)

Thus putting $x = 0, 1, 2, 3, 4$ successively in (*), we can determine the expected binomial frequencies as given below:

Number of Male births (x)	Expected frequency $f(x)$
0	$50\times{}^{4}C_{0}=50\times1=50$
1	$50\times{}^{4}C_{1}=50\times4=200$
2	$50\times{}^{4}C_{2}=50\times6=300$
3	$50\times{}^{4}C_{3}=50\times4=200$
4	$50\times{}^{4}C_{4}=50\times1=50$
Total	800

Let us now apply χ^2 test to examine the goodness of fit of the given data to the above binomial distribution. Calculations are given below:

	Observed frequency(O)	Expected frequency (E)	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
0	32	50	-18	324	6.48
1	178	200	-22	484	2.42
2	290	300	-10	100	0.33
3	236	200	36	1296	6.48
4	64	50	14	196	3.92
Total	800	800	0		19.63

Thus $\chi^2 = 19.63$

$\therefore$ The calculated value of χ^2 is 19.63

The degrees of freedom= $5-1=4$

The critical value of χ^2 at 5% level of significance for 4 degrees of freedom i.e., $\chi^2_{0.05,4} = 9.488$

Since the calculated value of χ^2 is greater than the critical value $\chi^2_{0.05,4}$, therefore we reject the null hypothesis and conclude that the hypothesis of equal male and female births is not valid.

Example 4: The figures given below are (a) the observed frequencies of a distribution, and (b) the frequencies of the normal distribution having the same mean, standard deviation and total frequency as in (a). Apply the χ^2 test of goodness of fit.

(a)	1	12	66	220	495	792	924	792	495	220	66	12	1
(b)	2	15	66	210	484	799	944	799	484	210	66	15	2

Solution: Since we are given the observed frequencies and expected frequencies of the normal distribution, therefore we apply the χ^2 test of goodness of fit.

Now we set up the null hypothesis that normal distribution is a good fit to the given frequency distribution, i.e., the given observed frequencies and the expected normal frequencies do not differ significantly.

In order to test the null hypothesis, the appropriate test statistic is given by

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where O is the observed frequency, E the expected frequency.

	O	E	$O-E$	$(O-E)^2$	$(O-E)^2/E$
	1, 12	2, 15	-4	16	0.94
	66	66	0	0	0.00
	220	210	10	100	0.48
	495	484	11	121	0.25
	792	799	-7	49	0.06
	924	944	-20	400	0.42
	792	799	-7	49	0.06
	495	484	11	121	0.25
	220	210	10	100	0.48
	66	66	0	0	0.00
	12, 1	15, 2	-4	16	0.94
Total	4096	4096	0		3.88

Thus $\chi^2 = 3.88$

$\therefore$ The calculated value of χ^2 is 3.88

The degrees of freedom=13-5=8

(The 13 original classes have been reduced to 11 by grouping, thus 2 degrees of freedom are lost in pooling. In addition, the mean, the standard deviation and total frequency of the original distribution have been used in calculating the theoretical frequencies, thus introducing three restraints. The number of degrees of freedom is accordingly 8.)

The critical value of χ^2 at 5% level of significance for 8 degrees of freedom i.e., $\chi^2_{0.05,8} = 15.507$

Since the calculated value of χ^2 is much less than the critical value $\chi^2_{0.05,8}$, it is highly non-significant. Hence, we may accept the null hypothesis and conclude that the normal distribution is a good fit to the given distribution.



CHECK YOUR PROGRESS

Q 2: In a particular chi-square goodness-of-fit test there are four categories and 200 observations. Use the 5% significance level.

(a) How many degrees of freedom are there?

(b) What is the critical value of chi-square?

Q 3: The null hypothesis and alternative hypothesis are

H_0 : The frequencies are equal and H_1 : The frequencies are not equal

Category	f
A	10
B	20
C	30

(a) State the decision rule ($\alpha = 0.05$), (b) Compute the value of Chi-square.

(c) What is your decision regarding

14.4.2 χ^2 Test for Independence of Attributes

In many business situations, a market researcher might be interested to study the relationship between two variables or to check whether they are independent of each other. For example, an edible oil company may be interested in knowing whether the purchase of oil is independent of the customer's age or whether it is dependent on the customer's age. These are two different situations and consequently the company needs to frame a production and selling strategy. Similarly following are the situations where the Chi-square test of independence can be employed to analyse attributes with multiple categories.

- Whether university students classified by gender are independent of courses of study.
- Whether wage-earners classified by education level are independent of income.
- Whether a product produced is acceptable or not is independent of the shifts in which it was produced.

Thus when observations are classified on the basis of two qualitative variables or attributes and arranged in a table, the resulting table is known as a contingency table (Table 14.1). χ^2 test of independence uses this contingency table for determining

independence of two qualitative variables.

Table 14.1: Contingency table

Variable B	Variable A				Total
	A_1	A_2	.....	A_k	
B_1	O_{11}	O_{12}	.....	O_{1k}	R_1
B_2	O_{21}	O_{22}		O_{2k}	R_2
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
B_r	O_{r1}	O_{r2}	.....	O_{rk}	R_r
Total	C_1	C_2		C_k	N

It is to be noted that the variables A and B have been classified into mutually exclusive categories. The value O_{ij} is the observed frequency for the cell in ith row and jth column. R_i is the total of the i^{th} row and C_j is the total of the j^{th} column. The row and column totals are added up to get a grand total (N), which represents the sample size. It is to be noted that to apply the Chi-square test we need to calculate the expected frequency.

The calculation of expected frequency, E_{ij} , corresponding to an observed frequency O_{ij} in row i and column j is based on the concept of multiplicative law of probability. Using this concept of probability the expected frequency, , is calculated as

$$E_{ij} = \frac{\text{i}^{\text{th}}\ \text{Row total}}{\text{sample size}} \times \frac{\text{j}^{\text{th}}\ \text{column total}}{\text{sample size}} \times \text{Grand total}$$

$$= \frac{\text{R}_\text{i}}{\text{N}} \times \frac{\text{C}_\text{j}}{\text{N}} \times \text{N} = \frac{\text{R}_\text{i} \times \text{C}_\text{j}}{\text{N}}$$

Thus expected frequency for any cell can be computed by using the formula

$$\text{E}_{\text{ij}} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

The analysis of a two-way contingency table helps to determine whether the two variables are unrelated or independent of each other. Thus the null hypothesis for a Chi-square test of independence is that the two variables are independent.

Under the null hypothesis, the χ^2 test statistic is defined as

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

The degrees of freedom for the Chi-square test of independence are given by

Degrees of freedom= (Number of rows-1) × (Number of columns-1)

$$= (r-1)\times(c-1)$$

Example 5: A certain drug is claimed to be effective in curing colds. In an experiment on 500 persons with cold, half of them were given the drug and half of them were given sugar pills. The patients' reactions to the treatment are recorded in the following table:

	Consequence			
Treatment	Helped	Reaction	No effect	Total
Drug	150	30	70	250
Sugar pills	130	40	80	250
Total	280	70	150	500

On the basis of the data, can it be concluded that there is a significant difference in the effect of the drug and sugar pills?

	Consequence			
Treatment	Helped	Reaction	No effect	Total
Drug	$\frac{250\times280}{500}=140$	$\frac{250\times70}{500}=35$	250-(140+O5)=75	250
Sugar pills	140	35	75	250
Total	280	70	150	500

Now we arrange the observed and expected frequencies in the following table in order to calculate the χ^2 test statistic.

Observed (O)	Expected (E)	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
150	140	10	100	0.714
130	140	-10	100	0.714
30	35	-5	25	0.714
40	35	5	25	0.714
70	75	-5	25	0.333
80	75	5	25	0.333
Total				3.522

Thus the χ^2 test statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 3.522$$

Number of degrees of freedom= (r-1) (c-1) = (2-1) (3-1) =2

The critical value of the χ^2 at 5% level of significance for 2 degrees of freedom = 5.99

Since the calculated value of the is less than the critical value, therefore we may accept the null hypothesis. Hence conclude that drug is no better than sugar pills in curing colds.

Example 6: Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students falling in different intelligence levels. The results are as follows:

Researcher	Number of students in each level				Total
	Below average	Average	Above Average	Genius	
X	86	60	44	10	200
Y	40	33	25	2	100
Total	126	93	69	12	300

Would you say that the sampling techniques adopted by the two researchers are significantly different?

Solution: We state he null hypothesis that the data obtained are independent of the sampling techniques adopted by the two researchers.

Under the null hypothesis of independence, the expected frequencies are calculated as:

$$E_{11} = \frac{126 \times 200}{300} = 84; \qquad E_{12} = \frac{93 \times 200}{300} = 62; \qquad E_{13} = \frac{69 \times 200}{300} = 46$$

Now we arrange the observed and expected frequencies in the following table in order to calculate the χ^2 test statistic.

Observed (O)	Expected (E)	$(O-E)$	$(O-E)^2$	$(O-E)^2/E$
86	84	2	4	0.048
60	62	-2	4	0.064
44	46	-2	4	0.087
10	8	2	4	0.500
40	42	-2	4	0.095
33	31	2	4	0.129
25	23	2	4	0.174
2	4	-2	4	1.000
Total	300	0		2.097

Thus the χ^2 test statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 2.097$$

Number of degrees of freedom= (r-1)(c-1)=(4-1)(2-1)=3

The critical value of the χ^2 at 5% level of significance for 3 degrees of freedom = 7.82

Since the calculated value of χ^2 the is less than the critical value, therefore we may accept the null hypothesis. Hence conclude that the sampling techniques adopted by the two investigators do not differ significantly.



CHECK YOUR PROGRESS

Q 4: A social scientist sampled 140 people and classified them according to income level and whether or not they played a state lottery in the last month. The sample information is reported below. Is it reasonable to conclude that playing the lottery is related to income level? Use the 5% significance level.

	Income			
	Low	Middle	High	Total
Played	46	28	21	95
Did not play	14	12	19	45
Total	60	40	40	140

(a) What is this table called?

(b) State the null and alternative hypotheses.

(c) What is the degrees of freedom?

(d) What is the decision rule?

(e) Determine the value of χ^2.

(f) Make a decision about H_0.

14.4.3 χ^2 Test for a Specified Population Variance

Let $x_1, x_2, ..., x_n$ be a random sample of size n drawn from a normal population with mean μ and variance σ^2. We want to test the hypothesis that the population variance has a specified value.

We set up the null hypothesis and the alternative hypothesis as

$$H_0 : \sigma^2 = \sigma_0^2 \text{ and } H_1 : \sigma^2 > \sigma_0^2$$

Then under the null hypothesis, the test statistic

$$\chi^2 = \sum \frac{\left(x - \bar{x}\right)^2}{\sigma^2} = \frac{ns^2}{\sigma^2}$$

follows χ^2 distribution with $(n-1)\ d.f.$, where

$s^2 = \frac{\sum\left(x - \bar{x}\right)^2}{n}$, denotes the sample variance.

By comparing the calculated value of χ^2 with the tabulated value of for at the specified level of significance, we may accept or reject the null hypothesis at that level of significance.

Remarks: 1. test given by

$$\chi^2 = \frac{ns^2}{\sigma^2}$$

can be applied only if the population is normal.

2. If the sample size is large, say, greater than 30, then we can use Fisher's approximation, viz.,

$$\sqrt{2\chi^2} \sim N\left(\sqrt{2n-1}, 1\right)$$

And then apply the normal test by considering the statistic

$$Z = \sqrt{2\chi^2} - \sqrt{2n-1} \sim N(0, 1)$$

Example 7: A battery manufacturer claims that the lives of his batteries have a standard deviation of 0.9 year. Can we accept manufacture's claim if the lives of a random sample of 10 batteries have a standard deviation of 1.2 years? (Use a 5% level of significance)

Solution: We set up the null and alternative hypotheses as

Null hypothesis $H_0 : \sigma = 0.9\ years$

Alternative hypothesis $H_1 : \sigma > 0.9\ years$

Under the null hypothesis, the test statistic

$\chi^2 = \frac{ns^2}{\sigma^2}$ ~Chi-square distribution with $(n-1)\ d.f.$

Now $\chi^2 = \frac{10 \times (1.2)^2}{(0.9)^2} = 17.78$

Degrees of freedom= n-1 =10-1=9

The critical value of χ^2 at 5% level for 9 d.f. i.e., $\chi^2_{0.05}(9) = 16.92.$

Since the critical value of χ^2 is greater than the tabulated value of value of χ^2 at 5% level for 9 d.f., we reject the null hypothesis and conclude that the standard deviation of life of the batteries is more than 0.9 years.

Example 8: The weights in kg. of 10 students are given below:

38, 40, 45, 53, 47, 43, 55, 48, 52, 49

Can we say that the variance of the distribution of weights of all students from which the above sample of 10 students was drawn, is equal to 20 square kg.?

Solution: We set up the null and alternative hypotheses as:

Null hypothesis $H_0 : \sigma^2 = 20\ sq.\ kg.$

Alternative hypothesis $H_1 : \sigma^2 > 20\ sq.\ kg.$

Under the null hypothesis, the test statistic

$\chi^2 = \frac{ns^2}{\sigma^2}$ ~Chi-square distribution with $(n-1)\ d.f.$

Now we prepare the following table to calculate the sample variance

Weight in kg. (x)	$(x-\bar{x})$ $x-47$	$(x-\bar{x})^2$
38	-9	81
40	-7	49
45	-2	4
53	6	36
47	0	0
43	-4	16
55	8	64
48	1	1
52	5	25
49	2	4
Total=470		280

$$\bar{x} = \frac{\sum x}{n} = \frac{470}{10} = 47 \qquad s^2 = \frac{\sum (x - \bar{x})^2}{n} = \frac{280}{10} = 28$$

Now, $\chi^2 = \frac{ns^2}{\sigma^2} = \frac{10 \times 28}{20} = 14$

Degrees of freedom= n-1=10-1=9

The critical value of χ^2 i.e., $\chi^2_{0.05}(9) = 16.92$

Since the calculated value of χ^2 is less than the critical value of $\chi^2_{0.05}(9)$, therefore we may accept our null hypothesis and conclude that the variance of the distribution of weights of all students in the population is 20 sq. kg.

14.4.4 χ^2 Test of Homogeneity

χ^2 test of homogeneity is used to determine whether two or more independent variables are drawn from the same population or from different populations. In other words, χ^2 test of homogeneity is used to know whether two or more populations are homogeneous with respect to some characteristics of interest. For example, a researcher may be interested in knowing whether the employees from three departments, production, finance, and personnel, feel the same about the requirements of the top management in terms of hard work expected from employees. The amount of hard work can also be classified into three groups, namely, very hard, hard, and easy going. In this situation, we can state a null hypothesis that the opinion of all the groups is the same about the requirement of hard work. In other words, the null hypothesis assumes that the three classifications are homogeneous in terms of their opinion about the amount of hard work required by the top management.

It may be noted that, in both χ^2 test of homogeneity and χ^2 test of independence, one is concerned with the cross tabulation of the data and hypotheses testing procedure is also same. The differences that exist between the χ^2 test of homogeneity and χ^2 test of independence are: (i) In χ^2 test of independence, a

researcher is interested in testing whether two attributes are independent. On the other hand in, χ^2 intest of homogeneity, a researcher investigates whether two or more populations are homogenous with respect to some characteristic of interest. (ii) in χ^2 test of homogeneity, two or more independent samples are drawn from each population whereas in case of χ^2 test of independence, a single sample is drawn from a population.

Example 9: A television company has launched a new product with some advanced features. The company wants to know the opinion of consumers about this product with respect to four characteristics: preferred brand with new features did not prefer brand with new features, preferred only a few new features, and indifferent. The company has divided consumers into three groups—executives/ officers; businessmen, and private consultants. It has taken a random sample of size 459 and obtained the following:

Consumers	Executives/Officers	Businessmen	Private consultants	Row total
Preferred brand with new features	35	25	40	100
Did not prefer brand with new features	30	45	34	109
Preferred only a few new features	45	50	25	120
Indifferent	25	55	50	130
Column total	135	175	149	459

Use χ^2 test of homogeneity and draw inference from the data.

Solution: Let us set up null and alternative hypotheses as

Null hypothesis H_0 : Opinion of all the groups is the same about the product with new features.

And alternative hypothesis H_1 : Opinion of all the groups is not the same about the product with new features.

In order to test the null hypothesis, the appropriate test statistic is given by:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where O is the observed frequency, E the expected frequency.

To calculate the expected frequencies for the corresponding observed frequencies, we apply the following formula

(Row total × Column total)/Grand total.

Now for the calculation of the expected frequencies against the corresponding observed frequencies and to compute the test statistic we prepare the following table:

Observed (O)	Expected (E)	$(O-E)$	$(O-E)^2/E$
35	29.41	5.59	1.0625
30	32.06	-2.06	0.1324
45	35.29	9.71	2.6717
25	38.24	-13.24	4.5841
25	38.13	-13.13	4.5213
45	41.56	3.44	0.2847
50	45.75	4.25	0.3948
55	49.56	5.44	0.5971
40	32.46	7.54	1.7514
34	35.38	-1.38	0.0538
25	38.95	-13.95	4.9962
50	42.20	7.8	1.4417
420			22.4917

Thus the χ^2 test statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 22.4917$$

The number of degrees of freedom $= (r-1)(c-1) = (4-1)(3-1) = 6$

The critical value of the χ^2 at 5% level of significance for 6 degrees of freedom = 12.59

Since the calculated value of the χ^2 is greater than the critical value, therefore we reject the null hypothesis and conclude that the opinion of all the groups is not the same about the product with new features. Hence, the company has to consider different groups and their needs separately.

Precautions about using the Chi-square test: In order to apply Chi-square hypothesis test properly, one has to be extremely careful and keep in mind certain precautions. First, a sample should

be large enough. If the expected frequencies are too small, the value of Chi-square gets over estimated. This will result in the rejection of the null hypothesis in several situations. To overcome this problem, we must ensure that the expected frequency in any cell of the contingency table should not be less than 5. In case the expected frequency less than 5 in more than one cell, we can combine these to obtain the expected frequency of at least 5. Another point to note is that the calculations must be made with actual numbers in each cell and not with proportions or percentages. If the proportions or percentages were used, then the theoretical distribution would not be applicable.

When the calculated value of Chi-square turns out to be more than the critical value at a specified level of significance, we reject our null hypothesis. On the other hand, when Chi-square value is less than the critical value, we may accept our null hypothesis. However, when the Chi-square value turns out to be zero, we have to be extremely careful to confirm that there is no difference between the observed and the expected frequencies. Such a situation may sometimes arise on account of a faulty manner used in the data collection.

In most of the cases, the problems of Chi-square involve simple calculations. However, for large sets of data the Chi-square test involves very comprehensive calculations. In all such situations, computer should be used. Several computer Statistics packages contain routines for carrying out Chi-square tests.



CHECK YOUR PROGRESS

Q 5: When chi-square test of homogeneity is useful?

……

……

……

……

……

14.5 TESTING OF HYPOTHESIS IN CASE OF SMALL SAMPLES (t-Test)

We have already discussed that in case of large samples (i.e. $n \geq 30$,) the sampling distribution of the statistic is approximately normal. But in many decision making situations, due to time, money and other constraints, sometimes a researcher may have to take small sample of size less than 30 (i.e. $n < 30$,). In case of small samples also, the researcher can apply standard normal test for a specified population mean or equality of two population means provided the sample is drawn from a normal population where S.D. σ is known. But if σ remains unknown then standard normal test fails and in all such cases the sampling distribution turns out to be the student's 't' distribution. Sir William Gosset of Ireland in early 1900, under his pen name 'Student', developed a method for testing of hypothesis known as 't-test'. It is to be noted that the concept of degrees of freedom is associated with t-test. The graph of the 't' distribution approaches the normal curve as the degrees of freedom increase.

If $x_1, x_2, \ldots\ldots\ldots\ldots, x_n$ be a random sample of size drawn from a normal population with mean μ and variance σ^2, then the Student's 't' statistic is defined by

$$t = \frac{\bar{x} - \mu}{S/\sqrt{n}},$$

where $\bar{x}$ = sample mean

and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$, sample mean square

is an unbiased estimator of σ^2. If s^2 be the sample variance given by

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

Then $ns^2 = (n-1)S^2$

Thus we have, $\frac{S}{\sqrt{n}} = \frac{s}{\sqrt{n-1}}$

$$\therefore \quad t = \frac{\bar{x} - \mu}{S/\sqrt{n}} = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$$

Here, 't' follows t-distribution with (n-1) d.f.

14.5.1 Characteristics of t-Distribution

The characteristics of t distribution are given below:

(1) It is an exact distribution.

(2) The distribution is unimodal with

Mean=Median=Mode=0

(3) The t-distribution is less peaked than normal distribution at the centre and higher at both the ends.

(4) The shape of the -distribution depends on depends on the sample size . As increases, it approaches the normal distribution.

(5) The limits of t is $-\infty < t < \infty$.

(6) The critical values of at level of significance vary with the degrees of freedom.

14.5.2 Applications of t-Distribution

The -distribution has a number of applications in Statistics some of which are given below

(i) It is used to test the significance of single mean, population standard deviation being unknown.

(ii) It is used to test the significant difference between two population means under the assumption that population standard deviations being equal but unknown.

(iii) It is used to test the significance of an observed sample correlation coefficient.

Assumptions for Student's -test: The following assumptions are made in Student's -test:

(i) The parent population from which the sample is drawn is normal.

(ii) The drawn sample is random.

(iii) The population standard deviation σ remains unknown.

14.5.3 Hypothesis Testing for Single Population Mean

Suppose a random sample of size $n(n<30)$ with mean $\bar{x}$ is drawn from a population μ. We want to test the hypothesis that the population has the specified value μ_0 when the population S.D. σ is unknown.

Here we set up the null hypothesis $H_0 : \mu = \mu_0$.

Now depending upon the situation, against this hypothesis we may have any one of the following three alternative hypotheses

(i) $H_1 : \mu \neq \mu_0$ (Two-tailed test)
(ii) $H_1 : \mu > \mu_0$ (Right-tailed test)
(iii) $H_1 : \mu < \mu_0$ (Left-tailed test)

Since the sample is small, therefore under the null hypothesis the test statistic to be constructed is

$t = \dfrac{x - \mu}{\dfrac{s}{\sqrt{n-1}}}$ follows t distribution with $\nu = (n-1)$ d f.

where $s =$ sample S.D., $\mu = \mu_0$, σ is the population standard deviation, n is the sample size and $\bar{x}$ is the sample mean.

Note: For large sample ($n \geq 30$), we have

S.E. ($\bar{x}$) = $\dfrac{s}{\sqrt{n}}$

On the other hand, for small sample ($n < 30$),

S.E. ($\bar{x}$) = $\dfrac{s}{\sqrt{n-1}}$

The critical values (i.e., tabulated values) of t distribution given in the table for 't' in the appendix are the critical values of for right-tailed test. The corresponding values for left-tailed test will just be the negative of these values. If the values of for a two-tailed test at level are to be determined then instead of α, we are to take $\alpha/2$.

If the alternative hypothesis is two-tailed as defined in (1), then the

null hypothesis is to be rejected if the calculated value of $|t| > t_{\alpha/2}$.

Again if the alternative hypothesis is one-tailed as defined in (2) or (3), the null hypothesis is to be rejected if the calculated value of $|t| > t_{\alpha}$.

Example 10: Royal Tyres has launched a new brand of tyres for tractors and claims that under normal circumstances the average life of the Tyres is 40,000 km. A retailer wants to test this claim and has taken a random sample of 8 Tyres that yields mean of 39,750 with a standard deviation of 2618.61. He tests the life of the Tyres under normal circumstance. (Use $\alpha = 0.05$)

Solution: Here the size of the sample $(i.e., n)$ is 8 (< 30) and the population standard deviation is unknown, so test can be applied to test the hypothesis.

Given $\bar{x} = 39750$ and $s = 2618.61$

Now, we set up the null and alternative hypotheses as:

Null hypothesis $H_0 : \mu = 40{,}000$

Alternative hypothesis $H_1 : \mu \neq 40{,}000$

Under the null hypothesis the test statistic is

$$t = \frac{x - \mu}{\frac{s}{\sqrt{n-1}}}$$

$$= \frac{39750 - 40000}{\frac{2618.61}{\sqrt{8-1}}}$$

$$= -\frac{250}{1070.677132} = -0.233$$

$$\therefore \quad |t| = |-0.233| = 0.233$$

Degrees of freedom= $n - 1 = 7$

The critical value of $|t_{0.05}(7)|$ for two-tailed test = 2.365

Since the critical value of $|t|$ is less than the critical value of $|t|$ at 5% level for 7 d.f., therefore we may accept the null hypothesis at 5% level of significance and conclude that the company's claim is valid under normal conditions.

Example 11: The mean price of shares of a company during the year 1992 was Rs. 85. In 1993, the mean price of shares of 20 randomly selected days is found to be Rs. 75 with a standard deviation of Rs. 10. Is there a significant difference in the prices of shares for the two years? (Given $|t_{0.05}(19)| = 2.093$)

Solution: Here the size of the sample $(i.e., n)$ is 20 (< 30) and the population standard deviation σ is unknown, so t test can be applied to test the hypothesis.

Given $\bar{x} = RS.75$ and $s = Rs.10$

Now, we set up the null and alternative hypotheses as:

Null hypothesis $H_0 : \mu = Rs.85$

Alternative hypothesis $H_0 : \mu \neq Rs.85$

Under the null hypothesis the test statistic is

$$t = \frac{x - \mu}{\frac{s}{\sqrt{n-1}}}$$

Which follows t distribution with $(n-1)$ d f.

Now $$t = \frac{75 - 85}{\frac{10}{\sqrt{20-1}}}$$

$$= \frac{-10}{\frac{10}{\sqrt{19}}} = -\sqrt{19} = -4.36$$

$$\therefore \quad |t| = |-4.36| = 4.36$$

Degrees of freedom = $20 - 1 = 19$

The critical value of $|t_{0.05}(19)|$ for two-tailed test = 2.093

Since the critical value of $|t|$ is greater than the critical value of $|t|$ at 5% level for 19 d.f., therefore we reject the null hypothesis at 5% level of significance and conclude that there is significant difference between the mean prices of shares in the two years.

Example 12: A fertilizer mixing machine is set to give 12 kg. of nitrate for every 100 kg. of fertilizer. Ten bags of 100 kg. each are examined. The percentage of nitrate so obtained is: 11, 14, 13, 12, 13, 12, 13, 14,11 and 12. Is there reason to believe that the machine is defective? (Use $\alpha = 0.05$)

Solution: Given $n = 10$

Let us consider the null and alternative hypotheses as

$H_0 : \mu = 12$ and $H_1 : \mu \neq 12$

We now calculate mean and standard deviation of the given sample as shown below:

Calculations of sample mean and standard deviation:

Variable x	$x - \bar{x} = x - 12.5$	$(x - \bar{x})^2$
11	-1.5	2.25
14	1.5	2.25
13	0.5	.25
12	-0.5	.25
13	0.5	.25
12	-0.5	.25
13	0.5	.25
14	1.5	2.25
11	-1.5	2.25
12	-0.5	.25
125	0	10.5

$\bar{x} = \frac{\sum x}{n} = \frac{125}{10} = 12.5$ and $s = \sqrt{\frac{1}{n}\sum (x - \bar{x})^2} = \sqrt{\frac{10.5}{10}} = \sqrt{1.05} = 1.025$

Under null hypothesis the test statistic is

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}}$$

Which follows t distribution with $(n-1)$ d.f

Now $$t = \frac{12.5 - 12}{\frac{1.025}{\sqrt{10-1}}}$$

$$= \frac{0.5}{\frac{1.025}{\sqrt{9}}} = \frac{0.5}{0.3417} = 1.46$$

$\therefore \quad |t| = |1.46| = 1.46$

Degrees of freedom= $10 - 1 = 9$

The critical value of $|t_{0.05}(19)|$ for two-tailed test = 2.262

Since the critical value of $|t|$ is less than the critical value of $|t|$ at 5% level for 9 d.f., therefore we may accept the null hypothesis at 5% level of significance and conclude that the manufacturer's claim is valid, i.e., the machine is not defective.

14.5.4 Hypothesis Testing for Difference between Two Population Means

In case of two populations, our concern is to test if two independent samples have been drawn from two normal populations having the same means, the population standard deviations being equal. Let $\overline{x_1}$ and $\overline{x_2}$ be the means of two independent random samples of sizes n_1 and n_2 ($n_1 < 30, n_2 < 30$) drawn from two normal populations having means μ_1 and μ_2 respectively with equal but unknown standard deviations σ_1 and σ_2.

Thus the null hypothesis is $H_0 : \mu_1 = \mu_2$, i.e., samples have been drawn from the normal populations with same means.

The alternative hypothesis H_1 may be any one of the following depending upon the situation.

(1) $H_1 : \mu_1 \neq \mu_2$ (two-tailed)

(2) $H_1 : \mu_1 < \mu_2$ (left-tailed)

(3) $.H_1 : \mu_1 > \mu_2$ (right-tailed)

Assuming the null hypothesis to be true, the test statistic t is given by

$$t = \frac{\overline{x_1} - \overline{x_2}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

i.e., the test statistic t follows t distribution with n_1+n_2-2 d.f

where $\overline{x_1} = \frac{1}{n}\sum x_1, \qquad \overline{x_2} = \frac{1}{n}\sum x_2$

and $S^2 = \frac{1}{n_1+n_2-2}\left[\sum\left(x_1 - \overline{x_1}\right)^2 + \sum\left(x_2 - \overline{x_2}\right)^2\right]$

$= \frac{1}{n_1+n_2-2}\left[n_1 s_1^2 + n_2 s_2^2\right]$, where $s_1^2 = \frac{1}{n_1}\sum\left(x_1 - \overline{x_1}\right)^2$ and $s_2^2 = \frac{1}{n_2}\sum\left(x_2 - \overline{x_2}\right)^2$

is an unbiased estimate of the common population variance σ^2 based on both the samples.

The decision to reject or accept the null hypothesis is similar to that used in one-sample tests.

Example 13: The IQs (intelligence quotients) of 16 students from one area of a city showed a mean of 107 with a standard deviation of 10, while the IQs of 14 students from another area of the city showed a mean of 112 with a standard deviation of 8. Is there a significant difference between the IQs of the two groups at 5% level of significance?

Solution:

Given $n_1 = 16, \quad n_2 = 14, \quad \overline{x_1} = 107, \quad \overline{x_2} = 112, \quad s_1 = 10$ and $s_2 = 8$

We set up null and alternative hypothesis as

Null hypothesis is $H_0 : \mu_1 = \mu_2$

Alternative hypothesis is $H_0 : \mu_1 \neq \mu_2$

Since sizes of both the samples are small and population standard deviations are unknown therefore to test the hypothesis we apply t test given by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

i.e., the test statistic t follows t distribution with n_1+n_2-2 d.f.

Now

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[n_1 s_1^2 + n_2 s_2^2\right]$$

$$= \frac{1}{16+14-2}\left[16(10)^2 + 14(8)^2\right]$$

$$= 89.142857$$

Thus $S = 9.44$

$$\therefore \quad t = \frac{112-107}{9.44\sqrt{\frac{1}{16} + \frac{1}{14}}} = 1.45$$

The degrees of freedom = $n_1+n_2-2=28$

The critical value of $|t_{0.05}(28)|$ for two-tailed test = 2.02

Since the critical value of $|t|$ is less than the critical value of $|t|$ at 5% level for 28 d.f., therefore we may accept the null hypothesis at 5% level of significance and conclude that the difference between the two population means is not significant.

Example 14: In a test given to two groups of students, the marks obtained are as given below:

First group : 25 32 30 32 24 14 32
Second group: 24 34 22 30 42 31 40 30 32 35

Examine the significance of the difference between the arithmetic mean of the marks secured by the students of the above two groups.

Solution: Here we set up null and alternative hypotheses as

Null hypothesis is $H_0 : \mu_1 = \mu_2$ i.e., there is no significant difference between mean marks secured by students of the two groups.

Alternative hypothesis is $H_0 : \mu_1 \neq \mu_2$

Since sizes of both the samples are small and population standard deviations are unknown therefore to test the hypothesis we apply t test given by

$$t = \frac{\overline{x_1} - \overline{x_2}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

i.e., the test statistic t follows t distribution with n_1+n_2-2 d.f.

Now

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[n_1 s_1^2 + n_2 s_2^2\right]$$

Calculations of $\overline{x_1}$, $\overline{x_2}$ and pooled sample standard deviations are given below:

Calculation of $\overline{x_1}$, $\overline{x_2}$ and S

x_1	$x_1 - \overline{x_1}$ $= x_1 - 27$	$(x_1 - \overline{x_1})^2$	x_2	$x_2 - \overline{x_2}$ $= x_2 - 32$	$(x_2 - \overline{x_2})^2$
25	-2	4	24	-8	64
32	5	25	34	2	4
30	3	9	22	-10	100
32	5	25	30	2	4
24	-3	9	42	10	100
14	-13	169	31	-1	1
32	5	25	40	8	64
			30	-2	4
			32	0	0
			35	3	9
189		266	320		350

Here $\overline{x_1} = \frac{\sum x_1}{n_1} = \frac{189}{7} = 27$ $\overline{x_2} = \frac{\sum x_2}{n_2} = \frac{320}{10} = 32$

Sample variance, $s_1^2 = \frac{1}{n_1}\sum\left(x_1 - \overline{x_1}\right)^2 = \frac{266}{7} = 38$

Sample variance, $s_2^2 = \frac{1}{n_2}\sum\left(x_2 - \overline{x_2}\right)^2 = \frac{350}{10} = 35$

Now, $S = \sqrt{\frac{7\times 38 + 10\times 35}{7+10-2}} = 6.41$

$$\therefore \quad t = \frac{27-32}{6.41\sqrt{\frac{1}{7}+\frac{1}{10}}}$$

$$= \frac{-5}{6.41\sqrt{\frac{17}{70}}} = -\frac{5}{6.41\times 0.49} = -1.592$$

$$\therefore \quad |t| = 1.592$$

The degrees of freedom = $n_1+n_2-2=15$

The critical value of $|t_{0.05}(15)|$ for two-tailed test = 2.131

Since the critical value of $|t|$ is less than the critical value of $|t|$ at 5% level for 28 d.f., therefore we may accept the null hypothesis at 5% level of significance and conclude that the mean marks obtained by the students of two groups do not differ significantly.

Example 15: Two salesmen, A and B, are employed by a company. To study if there is any significant difference between the average sales of the two salesmen, the HRD manager of the company conducted a survey yielding the following information:

	Salesman A	Salesman B
Number of sales	20	22
Average weekly sales (Rs. lakh)	30	25
Standard deviation (Rs lakh)	10	7

Solution: First we set up the null hypothesis that there is no significant difference between the average sales of the two salesmen. That is, $H_0 : \mu_1 = \mu_2$

The alternative hypothesis is $H_1 : \mu_1 \neq \mu_2$

Since sizes of both the samples are small and population standard deviations are unknown therefore to test the hypothesis we apply t test given by

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

i.e., the test statistic t follows t distribution with n_1+n_2-2 d.f.

Now

$$S^2 = \frac{1}{n_1 + n_2 - 2}\left[n_1 s_1^2 + n_2 s_2^2\right]$$

$$= \frac{1}{20 + 22 - 2}\left[20(10)^2 + 22(7)^2\right] = 76.95$$

Thus $S = 8.77$

$$\therefore \quad t = \frac{30 - 25}{8.77\sqrt{\frac{1}{20} + \frac{1}{22}}} = 1.84$$

The degrees of freedom = $n_1+n_2-2=40$

The critical value of $|t_{0.05}(40)|$ for two-tailed test = 2.021

Since the calculated value of $|t|$ is less than the critical value of $|t|$ at 5% level for 40 d.f., therefore we may accept the null hypothesis at 5% level of significance and conclude that the average sales by the two salesmen are not significantly different.



CHECK YOUR PROGRESS

Q 6: To the following information (a) State null and alternative hypotheses, (b) State the decision rule, (c) Compute the pooled estimate of the population variance, (d) Compute the test statistic. (e) What is your decision regarding H_0 ?

A random sample of 10 observations from one population revealed a sample mean of 23 and a S.D. of 4. A random sample of 8 observations from another population revealed a sample mean of

26 and a S.D. of 5. Use 5% level of significance to test if there is any significant difference between population means?

14.5.5 Paired t – Test for Difference between Two Population Means

In the t – test used for difference of means, the two samples were independent of each other whereas in case of paired t – test we have

(i) The sample sizes (n_1 and n_2) are equal.

(ii) The sample observations $(x_1, x_{2,}.....,x_n)$ and $(y_1, y_2,.....,y_n)$ are not completely independent but they are dependent in pairs i.e., the pairs $(x_1, y_1),(x_2, y_2),.....(x_n, y_n)$ correspond to the 1st, 2nd, ……,nth unit respectively.

For example, the management of a company plagued by poor productivity realizes the need to provide technical training to employees. It hires a researcher to measure the productivity levels of a sample of 25 employees. The productivity levels are measured again after a one-month technical training programme. In this kind of pre-and post-training study, which are selected before and after the study cannot be treated as independent because each observation in sample 1 is related to the observation in sample 2. The productivity scores obtained before training is related to the scores obtained after training because the two measurements are apply to the same person.

Now we explain the paired t – test for difference between the means of two related populations as follows:

Suppose a business concern is interested to know if advertisement is really effective in promoting sales of a particular product. Let $(x_1, x_{2,}.....,x_n)$ be a sample of sales of the product in departmental stores for a certain period before advertisement campaign and $(y_1, y_2,.....,y_n)$ be the sample of corresponding

sales of the same product and for period in the same departmental stores respectively. Now $(x_i, y_i), i = 1,2,.....,n$ is the pair of sales in the i[th] departmental store before and after advertisement. Suppose we want to test the significance of the difference between the sales, and.

Let $d_i = x_i - y_i, (i = 1,2,.....,n)$

We set up the null hypothesis that there is no significant change in sales after advertisement campaign. i.e., $H_0 : \mu_1 = \mu_2$.

The alternative hypothesis is $H_1 : \mu_1 \neq \mu_2$.

Under the null hypothesis, the test statistic is

$$t = \frac{\bar{d}}{S/\sqrt{n}} \sim t_{n-1}$$

where $\mathrm{d} = \mathrm{x} - \mathrm{y}, \quad \bar{\mathrm{d}} = \frac{1}{\mathrm{n}}\sum \mathrm{d}$

and $\mathrm{S}^2 = \frac{1}{\mathrm{n}-1}\sum\left(\mathrm{d} - \bar{\mathrm{d}}\right)^2 = \frac{1}{\mathrm{n}-1}\left[\sum \mathrm{d}^2 - \frac{\left(\sum \mathrm{d}\right)^2}{\mathrm{n}}\right]$

Example 16: An electronic company arranged a special training programme for one segment of its employees. In order to measure the change in the attitude of its employees after the training, the company has prepared a well-designed questionnaire consisting of 10 questions on a 1 to 5 rating scale (1 is strongly disagree and 5 is strongly agree). The company selected a random sample of 10 employees. The scores are given below:

Employees	Scores before training	Scores after training
1	25	32
2	26	30
3	28	32
4	22	34
5	20	32
6	30	28
7	22	25
8	20	30
9	21	25
10	24	28

Test whether there is a significant change in the attitude of employees after the training programme. (Use $\alpha = 0.10$)

Solution: Here, scores before training (x) and after training (y) are not independent but are paired together. Hence, to test the difference between the means of two related populations we apply the paired $t-$test.

We set up the null and alternative hypotheses as:

The null hypothesis $H_0 : \mu_1 = \mu_2$ i.e., there is no significant effect of the training.

The alternative hypothesis $H_1 : \mu_1 \neq \mu_2$.

Under the null hypothesis, the test statistic is

$$t = \frac{\bar{d}}{S/\sqrt{n}} \sim t_{n-1}$$

where $d = x - y, \quad \bar{d} = \frac{1}{n}\sum d$

and $$S^2 = \frac{1}{n-1}\sum\left(d - \bar{d}\right)^2 = \frac{1}{n-1}\left[\sum d^2 - \frac{\left(\sum d\right)^2}{n}\right]$$

Calculations of $\bar{d}$ and S^2 are given below:

Employees	Scores before training(x)	Scores after training(y)	$d = x - y$	d^2
1	25	32	-7	49
2	26	30	-4	16
3	28	32	-4	16
4	22	34	-12	144
5	20	32	-12	144
6	30	28	2	4
7	22	25	-3	9
8	20	30	-10	100
9	21	25	-4	16
10	24	28	-4	16
Total			-58	514

$$\therefore d = \frac{\sum d}{n} = \frac{-58}{10} = -5.8$$

$$S^2 = \frac{1}{10-1}\left[514 - \frac{(-58)^2}{10}\right] = \frac{1}{9}(514 - 336.4) = 19.733333$$

Thus $S = \sqrt{19.733333} = 4.44$

$$\therefore t = \frac{-5.8}{4.44/\sqrt{10}} = \frac{-5.8}{1.4047} = -4.13$$

$|t| = 4.13$

The degrees of freedom = 9

The critical value of $|t_{0.10}(9)|$ for two-tailed test = 1.833

Since the calculated value of $|t|$ is greater than the critical value of $|t|$ at 10% level for 9 d.f., therefore we reject the null hypothesis at 10% level of significance and conclude that there is a significant difference in the attitude of employees before and after the training.



CHECK YOUR PROGRESS

Q 7: To the following (a) State null hypothesis, (b) State the decision rule, (c) Compute the test statistic. (e) What is your decision regarding H_0 ?

An IQ test was administered to 5 persons before and after they were trained. The results are given below:

Candidate:	I	II	III	IV	V
IQ before training:	110	120	123	132	125
IQ after training:	120	118	125	136	121

Test if there is any change in IQ level after the training programme.

.....

.....

.....

.....

.....

.....



14.6 LET US SUM UP

- Statistical tests that do not require any prior information about the population are termed as non-parametric tests. This unit focuses on only χ^2 (Chi-square) distribution and the related χ^2 test. χ^2 distribution is the family of curves with each distribution being defined by the degree of freedom associated to it.

- χ^2 test can be used for a variety of purposes. χ^2 test provides a platform that can be used to ascertain whether the theoretical probability distribution coincides with the empirical sample distribution. This is commonly known as χ^2 goodness of fit test. It can also be used to test the independence of independence of two variables. χ^2 test of independence makes use of the contingency table to determine the independence of two variables. This test can also be used for estimating the population variance. Again the test of homogeneity is applied to determine whether two or more populations are homogeneous with respect to some characteristics of interest.

- The test can be applied for dependent samples and the testing procedure is termed as "paired t test or t test for related samples." Under this approach, observations in 1st sample are related to the observations in 2nd sample.



14.7 FURTHER READING

1) Sharma, J.K. (2007). Business Statistics. New Delhi. Pearson Education Ltd.
2) Hazarika, P.L. (2016). Essential Statistics For Economics And Business Studies. New Delhi. Akansha Publishing House.
3) Lind, D.A., Marshal, W.G., Wathen, S.A. (2009) Statistical Techniques in Business and Economics. New Delhi. Tata McGraw Hill Education Private Limited.

14.8 ANSWERS TO CHECK YOUR PROGRESS

Ans to Q No 1: (i) T, (ii) F, (iii) T, (iv) T, (v) T.

Ans to Q No 2: (a) , (b) The critical value of at 5% level of significance for 3 degrees of freedom i.e., $\chi^2_{0.05,3} = 7.8147$

Ans to Q No 3: $d.f. = 3-1 = 2$

The critical value of at 5% level of significance for 2 degrees of freedom i.e., $\chi^2_{0.05,2} = 5.9915$

(a) The decision rule is to reject if

(b) $\chi^2 = \frac{(10-20)^2}{20} + \frac{(20-20)^2}{20} + \frac{(30-20)^2}{20} = 10.0$

(c)Reject H_0.

Ans to Q No 4: (a) Contingency table.

(b) Null hypothesis is H_0 : There is no relationship between income and whether the person played the lottery.

Alternative hypothesis is H_1 : There is a relationship between income and whether the person played the lottery.

(c) The degrees of freedom, $d.f. = (2-1)\times(3-1) = 2$

(d) Reject H_0 if $\chi^2 > 5.9915$,The critical value of χ^2 at 5% level of significance for 2 degrees of freedom i.e., $\chi^2_{0.05,2} = 5.9915$

(e)

$$\chi^2 = \frac{(46-40.71)^2}{40.71} + \frac{(28-27.14)^2}{27.14} + \frac{(21-27.14)^2}{27.14} + \frac{(14-19.29)^2}{19.29} + \frac{(12-12.86)^2}{12.86} + \frac{(19-12.86)^2}{12.86}$$

= 6.544

(f) Reject H_0.

Ans to Q No 5: Chi-square test of homogeneity is useful in a situation when we intend to verify whether several populations are homogeneous with respect to some characteristic of interest.

Ans to Q No 6: (a) Null hypothesis is $H_0 : \mu_1 = \mu_2$, Alternative hypothesis is $H_1 : \mu_1 \neq \mu_2$,

(b) Here $d.f. = 10 + 8 - 2 = 16$, Therefore tabulated value of $|t_{0.05}(16)|$ for two-tailed test = 2.120

The decision rule is to reject H_0 if $|t| > 2.120$.

(c) Pooled estimate of the population variance, $S^2 = \frac{(10)4^2 + (8)5^2}{10 + 8 - 2} = 22.5$

(d) $t = \frac{23 - 26}{\sqrt{22.5\left(\frac{1}{10} + \frac{1}{8}\right)}} = -1.333$

(e) H_0 may be accepted.

Ans to Q No 7: (a) Null hypothesis is H_0: There is no change in IQ level after training programme.

(b) Reject H_0 if $t > 4.6$, the tabulated value of $|t_{0.05}(4)|$ for two-tailed test.

(c) $\bar{d} = \frac{10}{5} = 2, S = 5.477$ Therefore, $t = \frac{2}{5.477/\sqrt{5}} = 0.817$

(d) H_0 may be accepted.



14.9 MODEL QUESTIONS

Q 1: Define Chi-square distribution. Discuss the important properties of the distribution.

Q 2: Under what situations Chi-square test can be applied?

Q 3: Write down the applications of Chi-square test.

Q 4: The figures given below are (a) the theoretical frequencies of a distribution, and (b) the frequencies of the normal distribution having

the same mean, standard deviation and the total frequency as in (a).

(a)	1	5	20	28	42	22	15	5	2
(b)	1	6	18	25	40	25	18	6	1

Apply χ^2 -test of goodness of fit.

[Hints: Since the observed and the expected frequencies are less than 5 at the beginning and at end of the series, these frequencies are grouped

Degrees of freedom = 4, Tabulated value of $\chi^2_{0.05}(4) = 9.49$

H_0 may be accepted implying that the fit is good.]

Q 5: A Mumbai film director claims that his films are liked equally by males and females. An opinion survey of a random sample of 1000 film-goers revealed the following results:

	Liked	Disliked
Male	402	193
Female	245	160

Is the film director's claim supported by data?

[Given : $\chi^2_{0.05} =$ 3.84 5.99 7.81

d.f.= 1 2 3]

Q 6: A certain oil corporation conducts a seismic test on land to assess the possibility of extraction of oil from the land. From the past 200 records, the following test results and oil yields were obtained:

	Oil yield frequencies		
Test results	High	Medium	Low
Good	44	22	13
Fair	15	19	26
Bad	5	8	47

Can we say, at 1% level of significance that oil yields and test results are associated?

Q 7: The following table gives the classification of 100 workers according to sex and the nature of work. Test whether the nature of work is independent of the sex of the work.

	Skilled	Unskilled
Male	40	20
Female	10	30

Q 8: A certain drug was administered to 456 males out of a total of 720 in a certain locality to test its efficiency against typhoid. The incidence of typhoid is shown below. Find out the effectiveness of the drug against the disease. (The table value of χ^2 for 1 degree of freedom at 5% level of significance is 3.84)

	Infection	No infection	Total
Administering the drug	144	312	456
Without administering the drug	192	72	264
Total	336	384	720

Q 9: A sample of 20 observations gave a standard deviation of 3.72. Is this compatible with the hypothesis that the sample is drawn from a normal population with variance 4.35?

[Hints: H_0 : The sample is drawn from a normal population.
Under H_0, the test statistic $\chi^2 = \frac{ns^2}{\sigma^2}$ follows χ^2 distribution with $n-1 = 20-1 = 19\ d.f.$ Table value of $\chi^2_{0.05}(19) = 30.144.$ is rejected at 5% level.]

Q 10: The sales data of an item in six shops before and after a special promotional campaign are as under:

Shops :	A	B	C	D	E	F
Before campaign :	35	28	31	48	50	42
After campaign :	58	29	30	55	56	45

Can the campaign be judged to be a success? Test at 5% level of significance?

(Significant values of $|t|$ at 1% level and 5% level for 5 d.f. respectively are 2.015 and 2.571)



REFERENCES (For All Units of the Course)

1) Feller, W. (1968). *An Introduction to Probability Theory and its Application*, Vol I, Wiley Eastern.
2) Gujrati, D. (1995). *Basic Econometrics*. McGraw-Hill, Inc., Singapore
3) Gun, A. M., Gupta, M K., & Dasgupta, B. (2002). *Fundamentals of Statistics*, Volume One; World Press Private Ltd. Kolkata
4) Gupta, S. C. (1994). *Fundamentals of Statistics*. Himalayan Publishing House, New Delhi.
5) Gupta, S. P. (1995). *Statistical Methods*. Sultan Chand & Sons, New Delhi.
6) Hazarika, P.L. (2016). Essential Statistics For Economics And Business Studies. New Delhi. Akansha Publishing House.
7) Lind, D.A., Marshal, W.G., Wathen, S.A. (2009) Statistical Techniques in Business and Economics. New Delhi. Tata McGraw Hill Education Private Limited.
8) Medhi, J. (2000). *Statistical Methods: An Introductory Text*, New Age International (P) Limited, Publishers.
9) Mukhopadhyay, P. (1996), *Mathematical Statistics*. New Central Book Agency (P) Ltd.
10) Mukhopadhyay, P. (1991). *Theory of Probability*. New Central Book Agency (P) Ltd.
11) Sharma, J.K. (2007). Business Statistics. New Delhi. Pearson Education Ltd.

Name of the Paper: .....
Course Code: .....

Centre for Internal Quality Assurance (CIQA)
Krishna Kanta Handiqui State Open University
City Office: Housefed Complex, Guwahati-781006

Learner's Feedback on Course

Dear Learner,

Regarding the course as mentioned above, we would like to know your opinions and comments so as to improve the quality of self learning materials in future. Please respond to the following statements by ticking the number you feel most reflect your opinion. After completion of the additional comments, please detach the page and send/mail the same to us at the address given below.

The Director, Centre for Internal Quality Assurance, KKHandiqui State Open University
Housefed Complex, Dispur, Guwahati-781006
(E-mail id: ciqa@kkhsou.in)

1) Approximately how many hours did you spend for studying the units in the course? ☐☐☐

2) Please give your opinions (by ✓ mark) to the following items based on your reading of the block:

Sl. No.	Statements	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
I)	The SLMs of the course fulfil the learning objectives set out in the beginning of each unit					
II)	The units of the SLM could satisfy my academic needs and expectations					
III)	The Unit writers have excellent knowledge about the course contents					
IV)	Language and contents of the units were presented at a level which I could readily understand					
V)	Ample opportunity for participation in the activities provided in the units					
VI)	Used enough Illustrations (Diagrams, tables etc.) for conceptual clarity					
VII)	Quality of content is engaging, relevant, and up-to-date					
VIII)	The self check questions are very helpful					
IX)	The Possible/model questions and the answers to check my progress have benefited me a lot					

Additional Comments: *(Please feel free to provide your open comments)*

1) Which aspects of the SLM, according to you, worked well?

.....

.....

2) What sort of changes/improvements do you feel KKHSOU could implement to improve the overall quality of the SLM?

.....

.....

Thank you for taking the time to complete this form.